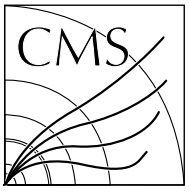


CERN-LHCC-2005-023
CMS TDR 7
20 June 2005

CMS

The Computing Project

Technical Design Report

CMS Computing Project		
CMS Spokesperson	Michel Della Negra, CERN	Michel.Della.Negra@cern.ch
CMS Technical Coordinator	Alain Herve, CERN	Alain.Herve@cern.ch
CMS Collab. Board Chair (also CPT Institution Board)	Lorenzo Foa, Pisa	Lorenzo.Foa@cern.ch
CPT Project Manager	Paraskevas Sphicas, CERN/Athens	Paraskevas.Sphicas@cern.ch
CPT Computing Coordinator	Lothar Bauerdick, FNAL	Bauerdick@fnal.gov
CPT Computing Coordinator	David Stickland, Princeton	David.Stickland@cern.ch
CPT Resource Manager	Lucas Taylor, Northeastern	Lucas.Taylor@cern.ch
CPT Architect	Vincenzo Innocente, CERN	Vincenzo.Innocente@cern.ch

Acknowledgements

The CMS Computing Group gratefully acknowledges the contributions of technical staff throughout the world who have been involved in the design, operations and analysis of the computing challenges that have led us to this TDR.

The CMS computing technical design has been developed in cooperation with our colleagues in the Worldwide LHC Computing Grid, together with the computing teams of ALICE, ATLAS and LHCb. We thank them all for their collaboration and for their assistance with and operation of the data-challenges that form the underpinning of this report.

We would like to thank Neil Geddes who has diligently followed the drafts of this TDR and contributed many important suggestions. We thank also the external reviewers of our Computing Model Paper: Neil, Tony Cass and John Harvey. We thank our CMS internal reviewers Jim Branson, Bob Clare, Lorenzo Foa and Gail Hanson for all their help and constructive comments.

For their perpetual good humour in the face of our unreasonable requests and deadlines, we thank the CMS Secretariat: Kirsti Aspola, Madeleine Azeglio, Nadejda Bogolioubova, Dorothée Denise, Dawn Hudson, Guy Martin and Marie-Claude Pelloux.

Special thanks to Sergio Cittolin for his artistic interpretation of the CMS Computing Model shown on the cover page

We also wish to thank our collaborators on CMS and especially the CMS management for their continuous support and encouragement.

ISBN 92-9083-252-5

Also available at: `http://cmsdoc.cern.ch/cms/cpt/tdr/`

CMS Collaboration

Yerevan Physics Institute, Yerevan, ARMENIA
G.L. Bayatian, S. Chatrchyan, A.M. Sirunyan, S. Stepanian

Institut für Hochenergiephysik der OeAW, Wien, AUSTRIA
W. Adam, T. Bergauer, J. Erö, M. Friedl, R. Fruehwirth, J. Hrubec, M. Jeitler, M. Krammer, I. Magrans, W. Mitaroff, N. Neumeister[**1], M. Pernicka, P. Porth, H. Rohringer, H. Sakulin, J. Strauss, A. Taurok, W. Waltenberger, G. Walzel, E. Widl, C.-E. Wulz

Research Institute for Nuclear Problems, Minsk, BELARUS
A. Fedorov, N. Gorodichenine, M. Korzhik, V. Panov, V. Yeudakimau, R. Zuyeuski

National Centre for Particle and High Energy Physics, Minsk, BELARUS
V. Chekhovsky, U. Chmel, I. Emeliantchik, M. Kryvamaz, A. Litomin, V. Mossolov, S. Reutovich, N. Shumeiko, A. Solin, A. Tikhonov, V. Zalessky

Byelorussian State University, Minsk, BELARUS
V. Petrov

Vrije Universiteit Brussel, Brussel, BELGIUM
J. D'Hondt, S. De Weirdt, R. Goorens, J. Heyninck, S. Lowette, S. Tavernier, W. Van Doninck[**2], L. Van Lancker

Université Libre de Bruxelles, Bruxelles, BELGIUM
O. Bouhali, B. Clerbaux, P. Marage, L. Neukermans, V. Sundararajan, C. Vander Velde, P. Vanlaer, J. Wickens

Université Catholique de Louvain, Louvain-la-Neuve, BELGIUM
S. Assouak, J.L. Bonnet, B. De Callatay, J. De Favereau De Jeneret, G. De Hemptinne, S. De Visscher, C. Delaere, P. Demin, D. Favart, E. Forton, G. Grégoire, T. Keutgen, G. Leibenguth, V. Lemaitre, Y. Liu, D. Michotte, O. Militaru, A. Ninane, S. Ovyn, K. Piotrzkowski, V. Roberfroid, X. Rouby, O. Van der Aa, M. Vander Donckt

Université de Mons-Hainaut, Mons, BELGIUM
E. Daubie, P. Herquet, A. Romeyer

Universiteit Antwerpen, Wilrijk, BELGIUM
W. Beaumont, E. De Langhe, E. De Wolf, M. Tasevsky

Centro Brasileiro de Pesquisas Fisicas, Rio de Janeiro, RJ, BRAZIL
M. Henrique Gomes E Souza

Instituto de Fisica Teorica-Universidade Estadual Paulista, Sao Paulo, SP, BRAZIL
S. Novaes

Universidade do Estado do Rio de Janeiro, Rio de Janeiro, RJ, BRAZIL
A. Santoro

Instituto de Fisica - Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, BRAZIL
J. Barreto, M. Vaz

Institute for Nuclear Research and Nuclear Energy, Sofia, BULGARIA
T. Anguelov, I. Atanasov, J. Damgov, N. Darmenov[**2], L. Dimitrov, V. Genchev[**2], P. Iaydjiev, B. Panev, S. Piperov, G. Sultanov, I. Vankov

University of Sofia, Sofia, BULGARIA
A. Dimitrov, V. Kozhuharov, L. Litov, M. Makariev, E. Marinova, M. Mateev, B. Pavlov, P. Petkov, C. Sabev, S. Stoynev, Z. Toteva[**2], V. Verguilov

Institute of High Energy Physics, Beijing, CHINA
J. Bai, J.G. Bian, G.M. Chen, H.S. Chen, Y.N. Guo, K. He, G. Huang, C.H. Jiang, Z.J. Ke, B. Li, J. Li, W.G. Li, H. Liu, G. Qin, J.F. Qiu, X.Y. Shen, G. Sun, H. Sun, C. Teng, Y.Y. Wang, Z. Xue, M. Yang, X. Yue, S.Q. Zhang, Y. Zhang, W. Zhao, G.Y. Zhu

Peking University, Beijing, CHINA
Y. Ban, J. Cai, K. Kang, S.J. Qian, Y.L. Ye, J. Ying

University for Science and Technology of China, Hefei, Anhui, CHINA
J. Wu, Z.P. Zhang

Shanghai Institute of Ceramics, Shanghai, CHINA (associated institute)
Q. Deng, P.J. Li, D.Z. Shen, Z.L. Xue, D.S. Yan, P. Yang[**2], H. Yuan

Technical University of Split, Split, CROATIA
N. Godinovic, I. Puljak, I. Soric

University of Split, Split, CROATIA
Z. Antunovic, M. Dzelalija, K. Marasovic

University of Cyprus, Nicosia, CYPRUS
C. Nicolaou, A. Papadakis, P.A. Razis, D. Tsiakkouri

National Institute of Chemical Physics and Biophysics, Tallinn, ESTONIA
A. Hektor, M. Kadastik, E. Lippmaa, M. Müntel, M. Raidal[**2]

Laboratory of Advanced Energy Systems, Helsinki University of Technology, Espoo, FINLAND
P.A. Aarnio

Helsinki Institute of Physics, Helsinki, FINLAND
S. Czellar[**2], E. Haeggstroem, M.A. Heikkinen, J. Härkönen, N. Jiganova, V. Karimäki[**2], R. Kinnunen, T. Lampén, K. Lassila-Perini, S. Lehti, T. Lindén, P.R. Luukka, T. Mäenpää, J. Nysten, E. Tuominen, J. Tuominiemi, D. Ungaro[**2], L. Wendland

Lappeenranta University of Technology, Lappeenranta, FINLAND
T. Tuuva

Laboratoire d'Annecy-le-Vieux de Physique des Particules, IN2P3-CNRS, Annecy-le-Vieux, FRANCE
J.P. Guillaud, P. Nedelec, D. Sillou

DSM/DAPNIA, CEA/Saclay, Gif-sur-Yvette, FRANCE
M. Anfreville, E. Bougamont, P. Bredy, R. Chipaux, M. Dejardin, D. Denegri, J. Descamps, B. Fabbro, J.L. Faure, F.X. Gentit, A. Givernaud, P. Gras, G. Hamel de Monchenault, P. Jarry, F. Kircher, M.C. Lemaire, B. Levesy[**2], E. Locci, J.P. Lottin, I. Mandjavidze, M. Mur, E. Pasquetto, A. Payn, J. Rander, J.M. Reymond, F. Rondeaux, A. Rosowsky, Z.H. Sun, P. Verrecchia

Laboratoire Leprince-Ringuet, Ecole Polytechnique, IN2P3-CNRS, Palaiseau, FRANCE
M. Anduze, S. Baffioni, M. Bercher, U. Berthon, S. Bimbot, J. Bourotte[**2], P. Busson, M. Cerutti, D. Chamont, C. Charlot, C. Collard, D. Decotigny, E. Delmeire, L. Dobrzynski, A.M. Gaillac, Y. Geerebaert, J. Gilly, M. Haguenauer, A. Karar, A. Mathieu, G. Milleret, P. Miné, P. Paganini, P. Poilleux, T. Romanteau, I. Semeniouk, Y. Sirois

Institut de Recherches Subatomiques, IN2P3-CNRS - ULP, LEPSI Strasbourg, UHA Mulhouse, Strasbourg, FRANCE
J.D. Berst[**3], R. Blaes[**4], J.M. Brom, G.L. Claus, F. Didierjean, F. Drouhin[**2], J.C. Fontaine[**4], U. Goerlach, P. Graehling, L. Gross, D. Huss, C. Illinger[**3], P. Juillot, A. Lounis, C. Maazouzi, S. Michal, R. Strub, T. Todorov[**2], P. Van Hove, D. Vintache

Institut de Physique Nucléaire de Lyon, IN2P3-CNRS, Univ. Lyon I, Villeurbanne, FRANCE
M. Ageron, G. Baulieu, M. Bedjidian, A. Bonnevaux, E. Chabanat, C. Combaret, D. Contardo, R. Della Negra, P. Depasse, M. Dupanloup, T. Dupasquier, H. El Mamouni, N. Estre, J. Fay, S. Gascon, N. Giraud, C. Girerd, R. Haroutounian, J.C. Ianigro, B. Ille, M. Lethuillier, N. Lumb, D. Mercier, L. Mirabito[**2], S. Perries, O. Ravat, B. Trocme

High Energy Physics Institute, Tbilisi State University, Tbilisi, GEORGIA
R. Kvatadze

Institute of Physics Academy of Science, Tbilisi, GEORGIA
V. Roinishvili

RWTH, I. Physikalisches Institut, Aachen, GERMANY
R. Adolphi, R. Brauer, W. Braunschweig, H. Esser, L. Feld, A. Heister, W. Karpinski, K. Klein, C. Kukulies, S. König, J. Olzem, A. Ostapchuk, D. Pandoulas, G. Pierschel, F. Raupach, S. Schael, G. Schwering, M. Thomas, M. Wlochal

RWTH, III. Physikalisches Institut A, Aachen, GERMANY
A. Adolf, M. Bontenackels, A. Böhm, H. Fesefeldt, T. Hebbeker, S. Hermann, G. Hilgers, K. Hoepfner, C. Hof, S. Kappler, D. Lanske, B. Philipps, H. Reithler, M. Sowa, H. Szczesny, M. Tonutti, O. Tsigenov

RWTH, III. Physikalisches Institut B, Aachen, GERMANY
F. Beissel, M. Duda, G. Flügge, T. Franke, K. Hangarter, D. Heydhausen, S. Kasselmann, T. Kress, A. Linn, J. Mnich, A. Nowack, M. Poettgens, O. Pooth, M. Weber

University of Hamburg, Hamburg, GERMANY
U. Holm, R. Klanner, U. Pein, P. Schleper, N. Schrim, G. Steinbrueck, M. Stoye, R. Van Staa, K. Wick

Institut für Experimentelle Kernphysik, Karlsruhe, GERMANY
P. Blüm, V. Buege, W. De Boer, G. Dirkes, M. Erdmann, M. Fahrer, M. Feindt, U. Felzmann, J. Fernandez Menendez, M. Frey, A. Furgeri, F. Hartmann, S. Heier, C. Jung, T. Müller, T. Ortega Gomez, C. Piasecki, G. Quast, K. Rabbertz, D. Schieferdecker, A. Schmidt, H.J. Simonis, A. Theel, T. Weiler, C. Weiser, J. Weng[**2], V. Zhukov[**5]

University of Athens, Athens, GREECE
G. Bruno, G. Karapostoli[**2], P. Katsas, P. Kreuzer, N. Marinelli, A. Panagiotou, C. Papadimitropoulos

Institute of Nuclear Physics "Demokritos", Attiki, GREECE
G. Anagnostou, M. Barone, T. Geralis, C. Kalfas, A. Kyriakis, S. Kyriazopoulou, D. Loukas, A. Markou, C. Markou, A. Zachariadou

University of Ioánnina, Ioánnina, GREECE
A. Asimidis, X. Aslanoglou, I. Evangelou, P. Kokkas, N. Manthos, I. Papadopoulos, G. Sidiropoulos, F.A. Triantis, P. Vichoudis[**2]

KFKI Research Institute for Particle and Nuclear Physics, Budapest, HUNGARY
G. Bencze[**2], L. Boldizsar, P. Hidas, A. Laszlo, G. Odor, F. Sikler, G. Vesztergombi, P. Zalan

Institute of Nuclear Research ATOMKI, Debrecen, HUNGARY
J. Molnar

Kossuth Lajos University, Debrecen, HUNGARY
G. Marian, P. Raics, Z. Szabo, Z. Szillasi, G. Zilizi

Panjab University, Chandigarh, INDIA
S.B. Beri, V. Bhatnagar, M. Kaur, R. Kaur, J.M. Kohli, J. Singh

University of Delhi, Delhi, INDIA
A. Bhardwaj, S. Chatterji, B. Choudhary, M. Jha, R.K. Shivpuri, A. Srivastava

Bhabha Atomic Research Centre, Mumbai, INDIA
S. Borkar, M. Dixit, M. Ghodgaonkar, S.K. Kataria, S.K. Lalwani, V. Mishra, A. Topkar

Tata Institute of Fundamental Research - EHEP, Mumbai, INDIA
T. Aziz, S. Banerjee[**2], S. Chendvankar, P.V. Deshpande, M. Guchait[**6], A. Gurtu, G. Majumder, K. Mazumdar, M.R. Patil, K. Sudhakar, S.C. Tonwar

Tata Institute of Fundamental Research - HECR, Mumbai, INDIA
B.S. Acharya, S. Banerjee, S. Bheesette, S. Dugad, S.D. Kalmani, V.R. Lakkireddi, N.K. Mondal, N. Panyam, P. Verma

Institute for Studies in Theoretical Physics & Mathematics (IPM), Tehran, IRAN
H. Arfaei, M. Mohammadi

University College Dublin, Dublin, IRELAND
M. Grunewald

Università di Bari, Politecnico di Bari e Sezione dell' INFN, Bari, ITALY
K. Abadjiev, M. Abbrescia, L. Barbone, E. Cavallo, A. Colaleo[**2], T. Coviello, D. Creanza, N. De Filippis, M. De Palma, G. Donvito, L. Fiore, D. Giordano, G. Iaselli, F. Loddo, G. Maggi, M. Maggi, N. Manna, M.S. Mennea, S. My, S. Natali, S. Nuzzo, G. Pugliese, V. Radicci, F. Romano, G. Selvaggi, L. Silvestris, P. Tempesta, R. Trentadue, G. Zito

Università di Bologna e Sezione dell' INFN, Bologna, ITALY
G. Abbiendi, A. Benvenuti, D. Bonacorsi, S. Braibant-Giacomelli, P. Capiluppi, F. Cavallo, C. Ciocca, I. D'Antone, G.M. Dallavalle, F. Fabbri, A. Fanfani, P. Giacomelli[**7], C. Grandi, M. Guerzoni, L. Guiducci, S. Marcellini, G. Masetti, A. Montanari, C. Montanari, F. Navarria, F. Odorici, A. Perrotta, A. Rossi, T. Rovelli, G. Siroli, R. Travaglini

Università di Catania e Sezione dell' INFN, Catania, ITALY
S. Albergo, M. Chiorboli, S. Costa, M. Galanti, G. Gatto Rotondo, F. Noto, R. Potenza, G. Russo, C. Sutera, A. Tricomi, C. Tuve

Università di Firenze e Sezione dell' INFN, Firenze, ITALY
A. Bocci, G. Ciraolo, V. Ciulli, C. Civinini, R. D'Alessandro, E. Focardi, A. Macchiolo, N. Magini, F. Manolescu, C. Marchettini, S. Mersi, M. Meschini, S. Paoletti, G. Parrini, R. Ranieri, M. Sani

Università di Genova e Sezione dell' INFN, Genova, ITALY
P. Fabbricatore, M. Fossa, R. Musenich, C. Pisoni

Laboratori Nazionali di Legnaro dell' INFN, Legnaro, ITALY (associated institute)
S. Badoer, L. Berti, M. Biasotto, E. Frizziero, U. Gastaldi, M. Gulmini[**2], F. Lelli, G. Maron, A. Petrucci, S. Squizzato, N. Toniolo, S. Traldi

Istituto Nazionale di Fisica Nucleare e Universita Degli Studi Milano-Bicocca, Milano, ITALY
A. De Min, F. Ferri, G. Franzoni, A. Ghezzi, P. Govoni, M. Malberti, P. Negri, M. Paganoni, A. Pullia, S. Ragazzi, N. Redaelli, C. Rovelli, R. Salerno, T. Tabarelli de Fatis, S. Vigano'

Istituto Nazionale di Fisica Nucleare de Napoli (INFN), Napoli, ITALY
G. Comunale, F. Fabozzi, P. Paolucci, D. Piccolo, C. Sciacca

Università di Padova e Sezione dell' INFN, Padova, ITALY
N. Bacchetta, M. Bellato, M. Benettoni, D. Bisello, E. Borsato, A. Candelori, P. Checchia, E. Conti, U. Dosselli, V. Drollinger, F. Fanzago, F. Gasparini, U. Gasparini, M. Giarin, P. Giubilato, F. Gonella, A. Kaminskiy, S. Karaevskii, V. Khomenkov, S. Lacaprara, I. Lippi, M. Loreti, O. Lytovchenko, M. Mazzucato, A.T. Meneguzzo, M. Michelotto, F. Montecassiano[**2], M. Nigro, M. Passaseo, M. Pegoraro, P. Ronchese, E. Torassa, S. Vanini, S. Ventura, M. Zanetti, P. Zotto, G. Zumerle

Università di Pavia e Sezione dell' INFN, Pavia, ITALY
G. Belli, U. Berzano, R. Guida, M.M. Necchi, S.P. Ratti, C. Riccardi, G. Sani, P. Torre, P. Vitulo

Università di Perugia e Sezione dell' INFN, Perugia, ITALY
F. Ambroglini, E. Babucci, D. Benedetti, G.M. Bilei[**2], B. Checcucci, L. Fanò, M. Giorgi, P. Lariccia, G. Mantovani, D. Passeri, M. Pioppi, P. Placidi, V. Postolache, D. Ricci, A. Santocchia, L. Servoli, D. Spiga

Università di Pisa, Scuola Normale Superiore e Sezione dell' INFN, Pisa, ITALY
P. Azzurri, G. Bagliesi, A. Bardi, A. Basti, J. Bernardini, T. Boccali, L. Borrello, F. Bosi, R. Castaldi, R. Cecchi, C. Cerri, R. Dell'Orso, S. Donati, F. Donno, S. Dutta, L. Foà, S. Galeotti, S. Gennai, A. Giammanco, A. Giassi, S. Giusti, G. Iannaccone, L. Latronico, F. Ligabue, T. Lomtadze, B. Mangano, M. Massa, A. Messineo, A. Moggi, F. Morsani, F. Palla, F. Palmonari, F. Raffaelli, A. Rizzi, G. Segneri, G. Sguazzoni, P. Spagnolo, F. Spinella, R. Tenchini, G. Tonelli, A. Venturi, P.G. Verdini, M. Vos

Università di Roma I e Sezione dell' INFN, Roma, ITALY
S. Baccaro[**8], L. Barone, A. Bartoloni, F. Cavallari, S. Costantini, I. Dafinei, M. Diemoz, C. Gargiulo, E. Longo, P. Meridiani, G. Organtini, S. Rahatlou

Università di Torino e Sezione dell' INFN, Torino, ITALY
N. Amapane, F. Bertolino, C. Biino[**2], N. Cartiglia, M. Cordero, M. Costa, D. Dattola[**2], L. Demaria, C. Mariotti, S. Maselli, E. Menichetti, P. Mereu, E. Migliore, V. Monaco, M.M. Obertino, N. Pastrone, A. Romero, R. Sacchi, A. Staiano, P.P. Trapani

Kyungpook National University, Daegu, KOREA
D. Han, K.H. Kwon, D.C. Son

Chonnam National University, Kwangju, KOREA
J.Y. Kim

Konkuk University, Seoul, KOREA
S.Y. Jung, J.T. Rhee

Korea University, Seoul, KOREA
S.K. Park

Seoul National University, Seoul, KOREA
S.B. Kim

Centro de Investigacion y de Estudios Avanzados del IPN, Mexico City, Distrito Federal, MEXICO
H. Castilla Valdez

University of Auckland, Auckland, NEW ZEALAND
R. Gray, D. Krofcheck

University of Canterbury, Christchurch, NEW ZEALAND
M. Billinghurst, P. Bones, P. Butler

National Centre for Physics, Quaid-I-Azam University, Islamabad, PAKISTAN
Z. Aftab, M. Ahmad, U. Ahmad, I. Ahmed, J. Alam Jan, M.I. Asghar, S. Asghar, M. Hafeez, H.R. Hoorani, M. Iftikhar, M.S. Khan, N. Qaiser, T. Solaija, S. Toor

National University of Sciences And Technology, Rawalpindi Cantt, PAKISTAN (associated institute)
A. Ali, A. Bashir, A.M. Jan, A. Kamal, M. Saeed, S. Tanwir, M.A. Zafar

Institute of Experimental Physics, Warsaw, POLAND
M. Bluj, K. Bunkowski, M. Cwiok, H. Czyrkowski, R. Dabrowski, W. Dominik, K. Doroba, J. Krolikowski, I. Kudla, M. Pietrusinski, W. Zabolotny[**9], J. Zalipska, P. Zych

Soltan Institute for Nuclear Studies, Warsaw, POLAND
R. Gokieli, L. Goscilo, M. Górski, K. Nawrocki, G. Wrochna, P. Zalewski

Warsaw University of Technology, Institute of Electronic Systems, Warsaw, POLAND (associated institute)
K. Pozniak, R. Romaniuk

Laboratório de Instrumentação e Física Experimental de Partículas, Lisboa, PORTUGAL
R. Alemany-Fernandez, C. Almeida, N. Almeida, P. Bordalo, R. Bugalho De Moura, J. Gomes, A. Jain, M. Kazana, N. Leonardo, S. Ramos, J. Rasteiro Da Silva, P.Q. Ribeiro, M. Santos, J. Semiao, I. Teixeira, J.P. Teixeira, J. Varela[**2], N. Vaz Cardoso

Joint Institute for Nuclear Research, Dubna, RUSSIA
I. Anissimov, K. Babich, D. Bardin, I. Belotelov, V. Elsha, Y. Ershov, I. Filozova, A. Golunov, I. Golutvin, N. Gorbounov, I. Gramenitski, V. Kalagin, A. Kamenev, V. Karjavin, S. Khabarov, V. Khabarov, Y. Kiryushin, V. Konoplyanikov, V. Korenkov, V. Ladygin, A. Lanev, V. Lysiakov, A. Malakhov, I. Melnitchenko, G. Meshcheryakov, V.V. Mitsyn, P. Moisenz, K. Moissenz, S. Movchan, E. Nikonov, D. Oleynik, V. Palichik, V. Perelygin, A. Petrosyan, E. Rogalev, V. Samsonov, M. Savina, R. Semenov, S. Sergeev, S. Shmatov, S. Shulha, V. Smirnov, D. Smolin, A. Tcheremoukhine, O. Teryaev, E. Tikhonenko, A. Vishnevskiy, A. Volodko, N. Zamiatin, A. Zarubin, P. Zarubin, E. Zubarev

Petersburg Nuclear Physics Institute, Gatchina (St Petersburg), RUSSIA
A. Atamantchouk, A. Baldychev, N. Bondar, A. Goliach, V. Golovtsov, D. Goulevich, Y. Ivanov, V. Kim, E. Kouznetsova, V. Kozlov, V. Lazarev, V. Lebedev, E. Lobatchev, G. Makarenkov, E. Orishchin, B. Razmyslovich, V. Sknar, I. Smirnov, S. Sobolev, V. Tarakanov, I. Tkach, L. Uvarov, G. Velichko, S. Volkov, A. Vorobyev, D. Yakorev

Institute for Nuclear Research, Moscow, RUSSIA
I. Andreev, P. Antipov, G.S. Atoyan, S. Gninenko, N. Golubev, E.V. Gushin, M. Kirsanov, A. Kovzelev, N. Krasnikov, V. Matveev, A. Pashenkov, V.E. Postoev, V. Shirinyants, V. Shmatkov, A. Solovey, L. Stepanova, A. Toropin

Institute for Theoretical and Experimental Physics, Moscow, RUSSIA
A. Arefiev, V. Gavrilov, N. Ilina, V. Kaftanov[**2], I. Kiselevich, V. Kolosov, M. Kossov[**2], A. Krokhotin, S. Kuleshov, N. Luzhetskiy[**2], A. Oulianov, S. Semenov, V. Stolin, A. Usik, V. Zakharov

P.N. Lebedev Physical Institute, Moscow, RUSSIA
A.M. Fomenko, N. Konovalova, V. Kozlov, A.I. Lebedev, N. Lvova, S. Potashov, S.V. Rusakov

Moscow State University, Moscow, RUSSIA
E. Boos, A. Ershov, R. Gloukhov, A. Gribushin, V. Ilyin, O.L. Kodolova[**2], I.P. Lokhtin, V. Mikhaylin, L. Sarycheva, V. Savrin, L. Shamardin, A. Snigirev, I. Vardanyan

High Temperature Technology Center of Research & Development Institute of Power Engineering (HTTC RDIPE), Moscow, RUSSIA (associated institute)
D. Chmelev, D. Druzhkin, A. Ivanov, V. Kudinov, O. Logatchev, S. Onishchenko, A. Orlov, V. Sakharov, V. Smetannikov, A. Tikhomirov, S. Zavodthikov

State Research Center of Russian Federation - Institute for High Energy Physics, Protvino, RUSSIA
V. Abramov, A. Annenkov[**10], I. Azhguirei, S. Belyanchenko, S. Bitioukov, P. Goncharov, S. Gordeev, V. Goussev, V. Grishin, A. Inyakin, V. Katchanov, A. Khmelnikov, E. Kolacheva, A. Korablev, Y. Korneev, A. Kostritski, A. Krinitsyn, V. Krychkine, E. Kvachina, O. Lapygina, A. Levine, I. Lobov, V. Lukanin, A. Markov, V. Medvedev, M. Oukhanov, V. Pak, V. Petrov, V. Pikalov, P. Podlesnyy, V. Potapov, A. Ryabov, A. Sannikov, Z. Simonova, E. Skvortsova, S. Slabospitski, A. Sobol, A. Soldatov, A. Sourkov[**2], S. Stepouchkine, A. Sytine, B. Tchuiko, S. Tereschenko, L. Tourtchanovitch, S. Troshin, N. Tyurin, A. Uzunian, A. Volkov, A. Zaitchenko, S. Zelepoukine[**11]

Electron National Research Institute, St Petersburg, RUSSIA (associated institute)
V. Lukyanov, G. Mamaeva, Z. Prilutskaya, I. Rumyantsev, S. Sokha, S. Tataurschikov, I. Vasilyev

Vinca Institute of Nuclear Sciences, Belgrade, SERBIA
P. Adzic, S. Drndarevic[**12], D. Maletic, P. Milenovic, J. Puzovic[**12], N. Smiljkovic[**2], M. Zupan

Centro de Investigaciones Energeticas Medioambientales y Tecnologicas, Madrid, SPAIN
M. Aguilar-Benitez, J. Alberdi, M. Aldaya Martin, P. Arce[**2], C. Burgos Lazaro, J. Caballero Bejar, E. Calvo, M. Cerrada, N. Colino, M. Daniel, B. De La Cruz, C. Fernandez Bedoya, A. Ferrando, M.C. Fouz, P. Garcia-Abia, J. Hernandez, M.I. Josa, J.M. Luque, J. Marin, A. Molinero, J.C. Oller, E. Perez Calle, L. Romero, J. Salicio, C. Villanueva Munoz, C. Willmott

Universidad Autónoma de Madrid, Madrid, SPAIN
C. Albajar, M. Fernandez, I. Jimenez, R. Macias[**2], R.F. Teixeira, J.F. de Trocóniz

Universidad de Oviedo, Oviedo, SPAIN
J. Cuevas, J.M. Lopez

Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, SPAIN
A. Calderon, D. Cano Fernandez, I. Diaz Merino, L.A. Garcia Moral, G. Gomez, I. Gonzalez, A. Lopez Virto, J. Marco, R. Marco, C. Martinez Rivero, F. Matorras, T. Rodrigo, D. Rodriguez Gonzalez, A. Ruiz Jimeno, M. Sobron Sanudo, I. Vila, R. Vilar Cortabitarte

CERN, European Organization for Nuclear Research, Geneva, SWITZERLAND
D. Abbaneo, S.M. Abbas, I. Ahmed, S. Akhtar, S. Ashby, P. Aspell, E. Auffray, M. Axer, A. Ball, N. Bangert, D. Barney, F. Beaudette, N. Bernardino Rodrigues, C. Bloch, P. Bloch, S. Bonacini, M. Bosteels, A. Branson, A.M. Brett, H. Breuker, V. Brigljevic[**13], O. Buchmuller, D. Campi, T. Camporesi, E. Cano, E. Carrone, A. Cattai, G. Cervelli, R. Chierici, J. Christiansen, T. Christiansen, S. Cittolin, E. Corrin, M. Corvo, S. Cucciarelli, B. Curé, G. Daskalakis, A. De Roeck, D. Delikaris, M. Della Negra, A. Dierlamm, A. Elliott-Peisert, M. Eppard, H. Foeth, R. Folch, S. Fratianni[**14], A. Frey, W. Funk, A. Gaddi, M. Gastal, J.C. Gayde, H. Gerwig, K. Gill, F. Glege, R. Gomez-Reino Garrido, J. Gutleber, M. Hansen, A. Hervé, A. Honma, M. Huhtinen, V. Innocente, W. Jank, P. Janot, C. Jones, K. Kloukinas, C. Lasseur, M. Lebeau, P. Lecoq, M. Letheren, C. Ljuslin, R. Loos, G. Magazzu, L. Malgeri, M. Mannelli, J.M. Maugain, F. Meijers, E. Meschi, F. Moortgat, A. Moutoussi, J. Nash, E. Noah Messomo, A. Oh, A. Onnela, M. Oriunno, L. Orsini, L. Pape, R. Paramatti, G. Passardi, A. Patino Revuelta, B. Perea Solano, G. Perinic, P. Petagna, A. Petrilli, A. Pfeiffer, M. Pimiä, R. Pintus, J.P. Porte, H. Postema, R. Principe, J. Puerta Pelayo, A. Racz, J. Rehn, S. Reynaud, P. Rodrigues Simoes Moreira, G. Rolandi, P. Rosinsky, D. Samyn, C. Schaefer, C. Schwick, P. Sempere Roldán[**15], A. Sharma, P. Sharp[**16], P. Siegrist, N. Sinanis, W. Snoeys, P. Sphicas[**17], M. Spiropulu, F. Szoncsó, O. Teller, N. Toth, D. Treille, J. Troska, M.H. Tsai, E. Tsesmelis, D. Tsirigkas, A. Tsirou, F. Vasey, L. Veillet, M. Weber, P. Wertelaers, M. Wilhelmsson, I.M. Willers, B. Wittmer

Paul Scherrer Institut, Villigen, SWITZERLAND
W. Bertl, K. Deiters, K. Gabathuler, S. Heising, R. Horisberger, Q. Ingram, D. Kotlinski, A. Macpherson[**2], D. Renker, T. Rohe

Institut für Teilchenphysik, Eidgenössische Technische Hochschule (ETH), Zürich, SWITZERLAND
B. Betev, P. Cannarsa[**2], G. Davatz, G. Dissertori, M. Dittmar, L. Djambazov, J. Ehlers, R. Eichler, W. Erdmann, G. Faber, K. Freudenreich, A.S. Giolo-Nicollerat, R. Goudard, C. Grab, A. Holzner, P. Ingenito, P. Lecomte, A. Lister, W. Lustermann, J.D. Maillefaud[**2], A. Nardulli, F. Nessi-Tedaldi, R.A. Ofierzynski, F. Pauss, U. Roser, H. Rykaczewski, F. Stoeckli, H. Suter, G. Viertel, H. Von Gunten

Universität Zürich, Zürich, SWITZERLAND
E. Alagoz, C. Amsler, V. Chiochia, A. Dorokhov, C. Hoermann, K. Prokofiev, H. Pruys, C. Regenfus, P. Robmann, T. Speer, S. Steiner

National Central University, Chung-Li, TAIWAN
S. Blyth, Y.H. Chang, E.A. Chen, A. Go, C.C. Hung, C.M. Kuo, W. Lin

National Taiwan University (NTU), Taipei, TAIWAN
P. Chang, Y. Chao, K.F. Chen, Z. Gao[**2], Y. Hsiung, J.G. Shiu, K. Ueno, Y. Velikzhanin, P. Yeh

Cukurova University, Adana, TURKEY
M. Bakirci, S. Cerci, I. Dumanoglu, S. Erturk, E. Eskut, S. Koylu, A. Kuzucu-Polatöz, H. Ozkurt, K. Sogut, G. Önengüt

Middle East Technical University, Physics Department, Ankara, TURKEY
A. Esendemir, H. Gamsizkan, C. Ozkan, S. Sekmen, M. Serin-Zeyrek, R. Sever, E. Yazgan, M. Zeyrek

Bogaziçi University, Department of Physics, Istanbul, TURKEY
K. Cankocak[**18], E. Gulmez, E. Isiksal[**19], M. Kaya[**20], S. Ozkorucuklu[**21]

Institute of Single Crystals of National Academy of Science, Kharkov, UKRAINE
B. Grinev, V. Senchyshyn

National Scientific Center, Kharkov Institute of Physics and Technology, Kharkov, UKRAINE
L. Levchuk, V. Popov, P. Sorokin

University of Bristol, Bristol, UNITED KINGDOM
D.S. Bailey, T. Barrass, J.J. Brooke, R. Croft, D. Cussans, R. Frazier, N. Grant, M. Hansen, G.P. Heath, H.F. Heath, B. Huckvale, C. Lynch, C.K. Mackay, S. Metson, D.M. Newbold, V.J. Smith, R.J. Tapper

Centre for Complex Cooperative Systems, University of the West of England, Bristol, UNITED KINGDOM (associated institute)
A. Anjum, N. Baker, F. Estrella[**2], R. McClatchey[**2], A. Solomonides

Rutherford Appleton Laboratory, Didcot, UNITED KINGDOM
S.A. Baird, K.W. Bell, R.M. Brown, D.J.A. Cockerill, J.A. Coughlan, P.S. Flower, V.B. Francis, M. French, J. Greenhalgh, R. Halsall, J. Hill, L. Jones, B.W. Kennedy, L. Lintern, A.B. Lodge, J. Maddox, Q. Morrissey, P. Murray, M. Pearson, S. Quinton, J. Salisbury, A. Shah, C. Shepherd-Themistocleous, B. Smith, M. Sproston, R. Stephenson, S. Taghavirad, I.R. Tomalin, J.H. Williams

Imperial College, University of London, London, UNITED KINGDOM
F. Arteche[**2], R. Bainbridge, G. Barber, P. Barrillon, R. Beuselinck, W. Bialas[**2], D. Britton, D. Colling, G. Dewhirst, S. Dris[**2], C. Foudas, J. Fulcher, G. Hall, G. Iles, P. Lewis, B.C. MacEvoy, O. Maroney, A. Nikitenko[**22], M. Noy, A. Papageorgiou, D.M. Raymond, M.J. Ryan, C. Seez, M. Takahashi, T. Virdee[**2], O. Zorba

Brunel University, Uxbridge, UNITED KINGDOM
C. Da Via, P.R. Hobson, P. Kyberd, J. Nebrensky, O. Sharif, L. Teodorescu[**23], S.J. Watts, I. Yaselli

Boston University, Boston, Massachusetts, USA
G. Antchev[**24], E. Hazen, A.H. Heering, D. Lazic, E. Machado, D. Osborne, J. Rohlf, L. Sulak, S. Wu

Brown University, Providence, Rhode Island, USA
D. Cutts, R. Hooper, G. Landsberg, R. Partridge

University of California, Davis, Davis, California, USA
R. Breedon, M. Case, M. Chertok, J. Conway, P.T. Cox, R. Erbacher, J. Gunion, B. Holbrook, W. Ko, R. Lander, D. Pellett, J. Smith, M. Tripathi, R. Vogt

University of California, Los Angeles, Los Angeles, California, USA
V. Andreev, K. Arisaka, D. Cline, R. Cousins, S. Erhan, M. Felcini, J. Hauser, M. Ignatenko, B. Lisowski, B. Liu, C. Matthey, J. Mumford, S. Otwinowski, Y. Pischalnikov, P. Schlein, Y. Shi, V. Valuev, M. Von Der Mey, R. Wallny, H.G. Wang, X. Yang, Y. Zheng

University of California, Riverside, Riverside, California, USA
R. Clare, D. Futyan[**2], J.W. Gary, M. Giunta[**2], G. Hanson, G. Pasztor[**25], B.C. Shen, V. Sytnik, D. Zer-Zion

University of California, San Diego, La Jolla, California, USA
S. Bhattacharya, J.G. Branson, J. Letts, T. Martin, M. Mojaver, H.P. Paar, M. Pieri, A. Rana, A. White, F. Würthwein

University of California, Santa Barbara, Santa Barbara, California, USA
A. Affolder, C. Campagnari, C. Hill, J. Incandela, D. White

California Institute of Technology, Pasadena, California, USA
D. Adamczyk, T. Azim[**26], A. Bornheim, J. Bunn, J. Chen, J.H. Choi, G. Denis, P. Galvez, M. Gataullin, E. Hughes, S. Iqbal, D. Lattka, T. Lee, I. Legrand, V. Litvine, D. Nae, H.B. Newman, C. Otey, S.P. Pappas, S. Ravot, S. Shevchenko, S. Singh, C. Steenberg, X. Su, J. Sucik, M. Thomas, F. Van Lingen, B.R. Voicu[**2], A. Weinstein, R. Wilkinson, X. Yang, L.Y. Zhang, K. Zhu, R.Y. Zhu

Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
T. Ferguson, M. Paulini, J. Russ, N. Terentyev, H. Vogel, I. Vorobiev

Fairfield University, Fairfield, Connecticut, USA
C.P. Beetz, G. Cirino, V. Podrasky, C. Sanzeni, D. Winn

Fermi National Accelerator Laboratory, Batavia, Illinois, USA
S. Abdullin[**22], M.A. Afaq[**2], M. Albrow, J. Amundson, G. Apollinari, M. Atac, J.A. Bakken, B. Baldin, L.A.T. Bauerdick, A. Baumbaugh, U. Baur, D.P. Eartly, J.E. Elias, V..D. Elvira, D. Evans, I. Fisk, J. Freeman, F.J..M. Geurts, G. Graham, D. Green, G.M. Guglielmo, Y. Guo, J. Hanlon, S. Hansen, R.M. Harris, S.L. Holm, E. James, M. Johnson, U. Joshi, S. Kossiakov, J. Kowalkowski, T. Kramer, E. La Vallie, M. Larwill, S. Los, L. Lueking, G. Lukhanin, S. Lusin[**2], K. Maeshima, S.J. Murray, V. O'Dell, M. Paterno, D. Petravick, R. Pordes, O. Prokofyev, V. Rasmislovich, N. Ratnikova, A. Ronzhin, V. Sekhri, E. Sexton-Kennedy, T. Shaw, R.P. Smith, L. Spiegel, M. Stavrianakou, I. Suzuki, W. Tanenbaum, S. Tkaczyk, R. Vidal, H. Wenzel, J. Whitmore, W.M. Wu, Y. Wu, A. Yagil, T. Yetkin[**27], J.C. Yun

University of Florida, Gainesville, Florida, USA
D. Acosta, P. Avery, V. Barashko, P. Bartalini, D. Bourilkov[**24], R. Cavanaugh, A. Drozdetski, R.D. Field, Y. Fu, L. Gray, D. Holmes, B.J. Kim, S. Klimenko, J. Konigsberg, A. Korytov, K. Kotov, P. Levchenko, A. Madorsky, K. Matchev, G. Mitselmakher, Y. Pakhotin, H. Pi, C. Prescott, L. Ramond, P. Ramond, J.L. Rodriguez, B. Scurlock, H. Stoeck, J. Yelton

Florida International University, Miami, Florida, USA
W. Boeglin, V. Gaultney, L. Kramer, S. Linn, P. Markowitz, B. Raue, J. Reinhold

Florida State University, Tallahassee, Florida, USA
M. Bertoldi, S. Hagopian, V. Hagopian, K.F. Johnson, J. Mc Donald, H. Prosper, J. Thomaston, H. Wahl

Florida Institute of Technology, Melbourne, Florida, USA
M. Baarmand, L. Baksay[**28], M. Hohlmann, I. Vodopianov

University of Illinois at Chicago (UIC), Chicago, Illinois, USA
M.R. Adams, R.R. Betts, E. Chabalina, C. Gerber, C. Smith

The University of Iowa, Iowa City, Iowa, USA
U. Akgun, A.S. Ayan, A. Cooper, P. Debbins, F. Duru, M. Fountain, N. George, E. McCliment, J.P. Merlo, A. Mestvirishvili, M.J. Miller, J.E. Norbeck, Y. Onel, I. Schmidt, S. Wang

Iowa State University, Ames, Iowa, USA
E.W. Anderson, O. Atramentov, J.M. Hauptman, J. Lamsa

Johns Hopkins University, Baltimore, Maryland, USA
B.A. Barnett, B. Blumenfeld, C.Y. Chien, D.W. Kim, P. Maksimovic, S. Spangler, M. Swartz

The University of Kansas, Lawrence, Kansas, USA
P. Baringer, A. Bean, D. Coppage, M. Murray

Kansas State University, Manhattan, Kansas, USA
T. Bolton, W.E. Kahl, F. Rizatdinova, R. Sidwell, N. Stanton, E. Von Toerne

University of Maryland, College Park, Maryland, USA
D. Baden, R. Bard, S.C. Eno, T. Grassi, N.J. Hadley, R.G. Kellogg, S. Kunori, A. Skuja

Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
R. Arcidiacono, M. Ballintijn, G. Bauer, C. Paus, S. Pavlon, C. Roland, G. Roland, K. Sumorok, S. Tether, B. Wyslouch

University of Minnesota, Minneapolis, Minnesota, USA
D. Bailleux, P. Cushman, A. De Benedetti, A. Dolgopolov, R. Egeland, W.J. Gilbert, J. Grahl, N. Pearson, R. Rusack, A. Singovsky

University of Mississippi, University, Mississippi, USA
L.M. Cremaldi, D. Sanders, D. Summers

University of Nebraska-Lincoln, Lincoln, Nebraska, USA
K. Bloom, D.R. Claes, A. Dominguez, C. Lundstedt, G.R. Snow

Northeastern University, Boston, Massachusetts, USA
G. Alverson, E. Barberis, I. Britvitch, G. Eulisse, A. Kuznetsov, Y. Musienko[**29], S. Muzaffar, I. Osborne, S. Reucroft, J. Swain, L. Taylor, L. Tuura

Northwestern University, Evanston, Illinois, USA
P. Gartung, B. Gobbi, M. Kubantsev, H. Schellman, E. Spencer, R. Tilden

University of Notre Dame, Notre Dame, Indiana, USA
B. Baumbaugh, N.M. Cason, M. Hildreth, D.J. Karmgard, A. Kharchilava, R. Ruchti, J. Warchol, M. Wayne

The Ohio State University, Columbus, Ohio, USA
B. Bylsma, L.S. Durkin, J. Gilmore, J. Gu, D. Herman, D. Larsen, T.Y. Ling, C.J. Rush, V. Sehgal

Princeton University, Princeton, New Jersey, USA
P. Elmer, W.C. Fisher, V. Gupta, J. Mans, D. Marlow, P. Piroué, D. Stickland, C. Tully, T. Wildish, S. Wynhoff, Z. Xie

Purdue University, West Lafayette, Indiana, USA
K. Arndt, K. Banicz, V.E. Barnes, G. Bolla, D. Bortoletto, A. Bujak, A.F. Garfinkel, O. Gonzalez Lopez, L. Gutay, N. Ippolito, Y. Kozhevnikov, A.T. Laasanen, V. Maroussov, D. Miller, J. Miyamoto, I. Pal, C. Rott, A. Roy, A. Sedov, I. Shipsey

Rice University, Houston, Texas, USA
G. Eppley, M. Matveev, T. Nussbaum, B.P. Padley, J. Roberts, A. Tumanov, P. Yepes

University of Rochester, Rochester, New York, USA
A. Bodek, H. Budd, Y.S. Chung, P. De Barbaro[**2], R. Demina, R. Eusebi, G. Ginther, Y. Gotra, E. Halkiadakis, A. Hocker, A. Khanov[**22], S. Korjenevski, W. Sakumoto, P. Slattery, P. Tipton

Rutgers, the State University of New Jersey, Piscataway, New Jersey, USA
E. Bartz, J. Doroshenko, P.F. Jacques, M.S. Kalelkar, L. Perera, R. Plano, S. Schnetzer, S. Somalwar, R. Stone, G. Thomson, T.L. Watts, S. Worm, A. lath

Texas Tech University, Lubbock, Texas, USA
N. Akchurin, K. Carrell, J. Cranshaw, K. Gumus, H. Kim, V. Papadimitriou, A. Sill, M. Spezziga, E. Washington, R. Wigmans, L. Zhang

University of Wisconsin, Madison, Wisconsin, USA
Y.W. Baek, D. Bradley, D. Carlsmith, P. Chumney, I. Crotty[**2], S. Dasu, F. Feyzi, T. Gorski, L. Greenler, M. Grothe, M. Jaworski, A. Lanaro, R. Loveless, W. Mason, D. Reeder, W.H. Smith, D. Wenman

Yale University, New Haven, Connecticut, USA
S. Dhawan, V. Issakov, H. Neal, A. Poblaguev, M.E. Zeller

Institute of Nuclear Physics of the Uzbekistan Academy of Sciences, Ulugbek, Tashkent, UZBEKISTAN
M. Belov, Y. Koblik, B.S. Yuldashev

**1: Also at Purdue University, West Lafayette, USA

**2: Also at CERN, European Organization for Nuclear Research, Geneva, Switzerland

**3: Also at Université Louis Pasteur, Strasbourg, France

**4: Also at Université de Haute-Alsace, Mulhouse, France

**5: Also at Moscow State University, Moscow, Russia

**6: Also at Tata Institute of Fundamental Research - HECR, Mumbai, India

**7: Also at University of California, Riverside, Riverside, USA

**8: Also at ENEA - Casaccia Research Center, S. Maria di Galeria, Italy

**9: Also at Institute of Electronic Systems, Technical University of Warsaw, Poland

**10: Also at Bogoroditsk Technical Plant, Moscow, Russia

**11: Also at Institut für Teilchenphysik, Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland

**12: Also at Faculty of Physics of University of Belgrade

**13: Also at Institute Rudjer Boskovic, Zagreb, Croatia

**14: Also at Politecnico di Torino, Torino, Italy

**15: Also at Universidad de Santiago de Compostela, Santiago de Compostela, Spain

**16: Also at Rutherford Appleton Laboratory, Didcot, United Kingdom

**17: Also at University of Athens, Athens, Greece

**18: Also at Mugla University, Turkey

**19: Also at Marmara University, Istanbul, Turkey

**20: Also at Kafkas University, Kars, Turkey

**21: Also at Suleyman Demirel University, Isparta, Turkey

**22: Also at Institute for Theoretical and Experimental Physics, Moscow, Russia

**23: Also at University of Bucharest, Bucuresti-Magurele, Romania

**24: Also at Institute for Nuclear Research and Nuclear Energy, Sofia, Bulgaria

**25: Also at KFKI Research Institute for Particle and Nuclear Physics, Budapest, Hungary
**26: Also at National University of Sciences And Technology, Rawalpindi Cantt, Pakistan
**27: Also at Cukurova University, Adana, Turkey
**28: Also at Kossuth Lajos University, Debrecen, Hungary
**29: Also at Institute for Nuclear Research, Moscow, Russia

Executive Summary

This document provides a top-level description of the organisation of the CMS Offline Computing systems.

This organisation relies upon a number of co-operating pieces:

- A tier-organised structure of computing resources, based on a Tier-0 centre at CERN and a small number of Tier-1 centres connected using high-speed networks.
- A relatively large number of Tier-2 analysis centres where physics analysis will be performed.
- A comprehensive and performant software framework designed for high-energy event streams.
- Workload management tools to coordinate work at the centres and data management tools to ensure the efficient use of computing resources and the integrity of the data, including adequate auditing and safekeeping of raw and reconstructed data, calibration data, and job parameters.
- A comprehensive project management plan so that the various project deliverables are tracked and potential bottlenecks are detected and eliminated.

These pieces are discussed in this document in terms of the CMS Data Model and the Physics Analysis Models. The workload management is considered in the context of integration into the LHC Computing Grid.

Structure of this document

Chapter 1, the Introduction, describes the context of this document.

Chapter 2 provides the motivation behind the top-level baseline CMS Computing Model and describes it in detail.

Chapter 3 describes the use of tiered computing centres within the Model.

Chapter 4 describes the computing services required to implement the Model.

Chapter 5 describes the formal project management plan and the resources required.

Finally, appendices are included to provide additional material. Appendix A lists the formal requirements and specifications for CMS offline computing. Appendix B describes references for further reading and the associated bibliography. Appendix C lists the current members of the CMS Computing Project. Appendix D is a glossary of abbreviations, acronyms and terms used in computing and CMS.

Contents

Figures

Tables

Chapter 1

Introduction

Computing and software are of paramount importance to the Compact Muon Solenoid (CMS) experiment. [1]

The LHC experiments have recognised the magnitude and complexity of the problem of computing in the LHC environment for almost ten years. This long lead time has allowed CMS to develop innovative software tools and services. We have also had the opportunity to develop the organisation in which that software must efficiently function. In addition we have learned to include project management in this mix to track and effectively manage both the progress towards our goals and the resources that are required to meet them.

The central fact of HEP computing that has arisen over this period is the Grid. CMS will be using Grid computing resources, services and toolkits as basic building blocks. Our Tier-0 computing centre at CERN will be directly connected to the experiment for initial processing and data archiving, but considerable processing will be done at remote Tier-1 centres, and analysis will be done at Tier-2 centres, all interconnected using Grid technology.

To make this system function smoothly and efficiently, we have had to learn to live in a Grid-enabled world. We have recently completed a redesign of our Computing Model to take into account the realities of the Grid. We have carefully evaluated our event data and organised it into streams that may be reconstructed, archived, replicated, split or skimmed, as required. We have examined requirements for auditing the data provenance and propagating it as the data travels from centre to centre and is possibly re-processed multiple times. We have performed a similar analysis for non-event data, such as calibrations and job parameter lists.

The Grid is more than just hardware and networks, of course. The people who operate and maintain those resources are part of organisations which must be integrated together. We are now developing the agreements that will cover the governance and responsibilities of centres, including the range of hardware and personnel expected for the various tiers.

We are developing additional services of our own that will couple the centres together

to create a cohesive system to enable the collaboration and individual collaborators to access and process the data. Data management services will be responsible for locating, storing and transferring the data in a safe, efficient and auditable manner. Workload management services will employ the data management services and additional (Grid and CMS) services to manage large computational tasks, such as re-processing a large data set, in a distributed environment.

We have created a project management plan to track our progress toward implementing these goals and keep the project on schedule. The plan as presented here includes a list of the main deliverables of the project and the project schedule and milestones.

The Software technical design will be separately described in the Physics TDR, but we note here that the software task includes the responsibility for delivering: (1) the core application software, (2) the software for physics and detector simulation, reconstruction, calibration and physics analysis, (3) the software to implement Higher Level Triggers and the associated algorithms, and (4) the software to assure the quality and integrity of CMS data.

We expect that the work outlined in this document will help lead to physics coming from CMS in a reliable and timely manner.

Chapter 2

Overview of the CMS Computing Model

2.1 Introduction

In this chapter we describe the physics motivations for the CMS Computing Model and give an overview of the Model itself. This chapter summarises and extends the key conclusions of the Computing Model paper [2].

2.2 Physics Overview and Context

This document is primarily concerned with the preparations for the first full year of LHC running, expected to be 2008. This first year will likely be characterised by a poorly understood detector, unpredictable machine performance, possibly inadequate computing infrastructure but also with the potential for significant physics discoveries. We expect to reprocess data often and must be able to make that data in its complexity and richness available to the collaboration so that their expertise can be brought to bear on detector, software, calibration and physics as effectively as possible. We will need good mechanisms to allow the data to be processed according to the priorities (be they detector understanding or Higgs searches). We will need to use all the Tiers of computing resources as effectively as possible, pre-locating data where they can be most efficiently processed and ensuring that the granularity of job queues at the sites is sufficient to steer the majority of computing resources to the experiments priorities, while also ensuring that individual physicists still have the possibility to explore original ideas.

These principles lead us to a baseline solution that emphasises:

- Fast reconstruction code (Frequent re-reconstruction)

- Streamed Primary Datasets (Priority driven distribution and processing)
- Distribution of Raw and Reconstructed data together (Easy access to raw detector information)
- Compact data formats (Multiple copies at multiple sites)
- Effective and efficient production reprocessing and bookkeeping systems (CMS physicists able to make full use of the system)

CMS will operate a structured analysis environment with analysis groups focusing on the main physics activities. We will define priorities on the activities at the Tier-0 and Tier-1 facilities predicated on satisfying the analysis group requirements. Particularly at start-up the limited resources must be used carefully and much of this Computing Model is designed to enable this prioritisation to be effectively implemented.

To motivate the computing planning in this document we have used an operations scenario as described in table 2.1. The Computing Model is insensitive to small changes in the luminosity profile as trigger thresholds will be adjusted up and down to maintain steady data rates as the running conditions vary.

Year	pp operations		Heavy Ion operations	
	Beam time (seconds/year)	Luminosity ($cm^{-2}s^{-1}$)	Beam time (seconds/year)	Luminosity ($cm^{-2}s^{-1}$)
2007	$2-3\times10^{6}$	$2-10\times10^{32}$	–	–
2008	10^{7}	2×10^{33}	10^{6}	5×10^{26}
2009	10^{7}	2×10^{33}	10^{6}	5×10^{26}
2010	10^{7}	10^{34}	10^{6}	5×10^{26}

Table 2.1: Scenario of LHC operation assumed for the purposes of this document.

2.3 Computing Overview and Context

The CMS Computing Model makes use of the hierarchy of computing Tiers as has been proposed in the MONARC [3] working group and in the First Review of LHC Computing [4]. The service agreements for such a hierarchy have been established in the LCG Memorandum of Understanding, and we do not re-discuss them here, although they form an under-pinning of our Computing Model.

We expect this ensemble of resources to form the Worldwide LHC Computing Grid. We use the term WLCG to define the full computing available to the LHC (CMS) rather than to describe one specific middleware implementation and/or one specific deployed Grid. We

expect to actually operate in a heterogeneous Grid environment but we expect the details of local Grid implementations to be largely invisible to CMS physicists (These Grids are described elsewhere, e.g.: LCG-2 Operations [5]; Grid-3 Operations [6]; EGEE [7]; NorduGrid [8]; Open Science Grid [9].

The WLCG and the regions bringing grid resources to CMS will be responsible for assuring a homogeneous interface to their varied Grid environments. The CMS computing project itself will be responsible for building application layers that can operate on a few, at most, well defined grid interfaces. In the following we assume that all resources are usable in this way, with finite and reasonable development and support required from CMS itself. CMS does not plan to devote any resources to making non-standard environments operational for CMS.

CMS has chosen to adopt a distributed model for all computing including the serving and archiving of the raw and reconstructed data. This assigns to some regional computing centres some obligations for safeguarding and serving portions of the dataset that in earlier experiments have been associated with the host laboratory. The CMS Computing Model includes a Tier-0 centre at CERN, a CMS Analysis Facility at CERN, several Tier-1 centres located at large regional computing centres, and many Tier-2 centres.

2.4 Data Flow Overview

The CMS DAQ system writes DAQ-RAW events (1.5 MB) to the High Level Trigger (HLT) farm input buffer. The HLT farm writes RAW events (1.5 MB) at a rate of 150 Hz. RAW events are classified in $\mathcal{O}(50)$ primary datasets depending on their trigger history (with a predicted overlap of less than 10%). Primary dataset definition is immutable. An additional express-line is also written with events that will be reconstructed with high priority. Primary datasets are grouped into $\mathcal{O}(10)$ online streams in order to optimise their transfer to the offline farm and the following reconstruction process. Data transfer from HLT to the Tier-0 farm must happen in real-time at a rate of 225 MB/s.

Heavy-Ion data at the same total rate (225 MB/s) will be partially processed in real-time on the Tier-0 farm. Full processing of the Heavy-ion data is expected to occupy the Tier-0 during much of the LHC downtime (between annual LHC pp running periods).

The first event reconstruction is performed without delay [1] on the Tier-0 farm which writes RECO events (0.25 MB). RAW and RECO versions of each primary dataset are archived on the Tier-0 MSS, which takes custodial responsibility for the first copy, a copy is transferred to a Tier-1 which takes custodial responsibility for this. Transfer to other Tier-1 centres is subject to additional bandwidth being available. Thus RAW and

[1] We also consider the possibility of some short delay that may be required to allow the use of updated calibrations. We anticipate using the CMS-CAF with access to the express streams to perform these calibrations with a maximum latency of order 24 hours

RECO are available either in the Tier-0 archive or in at least one Tier-1 centre. The first version of the Analysis Object Data (AOD, 0.05 MB) which are derived from RECO events and contain a copy of all the high-level physics objects plus a summary of other RECO information sufficient to support typical analysis actions (for example re-evaluation of calorimeter cluster positions or track refitting, but not pattern recognition). will also be produced in the Tier-0 reconstruction step and distributed to the Tier-1 centres (One full copy at each Tier-1)

The Tier-1 centres produce subsequent AOD versions, and distribute these new versions between themselves. Additional processing (skimming) of RAW, RECO and AOD data at the Tier-1 centres will be triggered by Physics Groups requests and will produce second and third (etc.) generation versions of the AOD as well as TAGS (0.01 MB) which contain high level physics objects and pointers to events (e.g., run and event number) and which allow their rapid identification for further study.

Tier-1 centres are responsible for bulk re-processing of RAW data, which is foreseen to happen up to three times per year at the start of LHC, reducing in future years as data quantity grows and algorithms mature.

Selected skimmed data, all AOD of selected primary streams, and a fraction of RECO and RAW events are transferred to Tier-2 centres which support iterative analysis of authorised groups of users. Grouping is expected to be done not only on a geographical but also on a logical basis, e.g. supporting physicists performing the same analysis or the same detector studies.

The CMS computing system is geographically distributed. Data are spread over a number of centres following the physical criteria given by their classification into primary datasets. Replication of data is given more by the need of optimising the access of most commonly accessed data than by the need to have data "close to home".

2.5 Event Model

2.5.1 Data Tiers

CMS will use a number of event data formats with varying degrees of detail, size, and refinement. Starting from the raw data produced from the online system successive degrees of processing refine this data, apply calibrations and create higher level physics objects.

Table 2.2 describes the various CMS event formats. It is important to note that, in line with the primary focus of this document, this table corresponds to the LHC startup period and assumes a canonical luminosity of $\mathcal{L} = 2 \times 10^{33}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. At this time the detector performance will not yet be well understood, therefore the event sizes are larger to accommodate looser thresholds and avoid rejection of data before it has been adequately

understood. The determinations of the data volume include the effects of re-processing steps with updated calibrations and application of new software and the copying of data for security and performance reasons.

2.5.2 Raw Data (RAW)

2.5.2.1 RAW Event Content and Size

Efforts to estimate occupancies for various sub-detectors are an ongoing effort within CMS. They impact not only detector design but obviously also the computing model and its budget. In the following we report numbers derived for the low luminosity running period ($2 \times 10^{33}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$), with a stable well understood detector. There may be multiple ways to measure event size with different formats, packing and compression schemes. The basic format used will be the one generated by the event builder as it assembles the data from the FED's and creates the input to the HLT farm. This will be designated DAQ-RAW.

The online HLT system will create "RAW" data events containing: the detector data, the L1 trigger result, the result of the HLT selections ("HLT trigger bits"), and some of the higher-level objects created during HLT processing.

The largest contributor is expected to be the silicon strip detector, and its projected size is 130 kB/event [10]. This number was derived using the latest tunes for the PYTHIA event generator and the full simulation of the CMS detector, and therefore reflects the current understanding of the experiment. Based on this and similar work in the other sub-detectors, an overall size estimate for the DAQ-RAW format, at an instantaneous luminosity of $2\times10^{33}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, of 300 kB/event is obtained. This represents a sort of ideal, minimum, event-size for the low-intensity running.

There are various reasons to expect that the event size in reality will be larger than this estimate and we identify the following factors:

- F_{Det} reflects the effects of adverse startup conditions, detector commissioning, not completely effective "zero-suppression";
- F_{HLT} reflects the need to commission and understand the HLT algorithms, must keep all intermediate results;
- F_{MC} reflects the MC being overly optimistic (may be the event generator or the detector simulation or quite likely both).

The first two are the hardest to estimate. The duration of their impact is as hard to predict as their scope. For the CDF experiment at the Tevatron Run II, F_{Det} was as large as 2.5 for a few months and F_{HLT} was 1.25 and lasted about a year. This represents experiment-specific "failure modes" and should be taken as such. As part of the following

Event Format	Content	Purpose	Event size (MByte)	Events / year	Data volume (PByte)
DAQ-RAW	Detector data in FED format and the L1 trigger result.	Primary record of physics event. Input to online HLT	1-1.5	1.5×10^9 $= 10^7$ seconds $\times$ 150Hz	–
RAW	Detector data after online formatting, the L1 trigger result, the result of the HLT selections ("HLT trigger bits"), potentially some of the higher-level quantities calculated during HLT processing.	Input to Tier-0 reconstruction. Primary archive of events at CERN.	1.5	3.3×10^9 $= 1.5 \times 10^9$ DAQ events $\times 1.1$ (dataset overlaps) $\times 2$ (copies)	5.0
RECO	Reconstructed objects (tracks, vertices, jets, electrons, muons, etc.) and reconstructed hits/clusters	Output of Tier-0 reconstruction and subsequent re-reconstruction passes. Supports re-finding of tracks, etc.	0.25	8.3×10^9 $= 1.5 \times 10^9$ DAQ events $\times 1.1$ (dataset overlaps) $\times$ [2 (copies of 1st pass) + 3 (reprocessings/year)]	2.1
AOD	Reconstructed objects (tracks, vertices, jets, electrons, muons, etc.). Possible small quantities of very localised hit information.	Physics analysis, limited refitting of tracks and clusters	0.05	53×10^9 $= 1.5 \times 10^9$ DAQ events $\times 1.1$ (dataset overlaps) $\times 4$ (versions/year) $\times 8$ (copies per Tier $-$ 1)	2.6
TAG	Run/event number, high-level physics objects, e.g. used to index events.	Rapid identification of events for further study (event directory).	0.01	–	–
FEVT	Term used to refer to RAW+RECO together (not a distinct format).		–	–	–

Table 2.2: CMS event formats at LHC startup, assuming a luminosity of $\mathcal{L} = 2 \times 10^{33}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The sample sizes (events per year) allow for event replication (for performance reasons) and multiple versions (from re-reconstruction passes).

exercise we use these as our central value estimators for CMS. The third one, F_{MC}, is easier to estimate. Using the CDF data and MC, the occupancy predicted by the MC is compared to that observed in data. The MC is underestimating the observed occupancy by a factor of 1.6. There is no obvious reason to expect CMS MC to get better results.

As discussed in the Computing Model [2] the RAW event size *at start-up* is estimated to be $\simeq 1.5\,\mathrm{MB}$, assuming a luminosity of $\mathcal{L} = 2 \times 10^{33}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$.

It should be clear that in making this estimation we have made a number of best-guesses based on experience at running experiments operating in similar conditions. It is unsafe to predict now when the various contingency factors can be decreased; nor can we know now how much worse the actual running conditions may be. This value of 1.5 MB event (entering the offline system) is then a best estimated central value; we cannot exclude that it will in fact be anywhere in the range 1-2 MB for the running in the first sustained LHC data-taking period.

The RAW event size in the third year of running is estimated to be $\simeq 1.0\,\mathrm{MB}$, assuming a luminosity of $\mathcal{L} = 10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. This asymptotic value accounts for two effects. As the luminosity increases so will the event size, due to the increase in the number of pile-up events. However, as the detector and machine conditions stabilise with time and become better understood, the event size, for a given luminosity, will decrease. The error on the quoted value is dominated by the uncertainties in the time evolution of the various F factors described above.

CMS therefore plans the computing requirements based on a full year running period in 2008 with event sizes of about 1.5 MB. Just-in-time purchasing of some media, but not of operational throughput capacity, could be considered. Actual requirements for the following years will in any case be re-evaluated in time for a sensible purchasing profiles.

2.5.2.2 RAW Event Rates

The physics reach of CMS is in practice likely to be limited by the available resources for triggering and/or event processing. Significant increases of the scope or scale of the trigger and of the computing resources could significantly increase the opportunities to explore more physics.

Since the early planning of the LHC experiments, the rate of events to permanent storage was cited as $\sim 10^2\,\mathrm{Hz}$. The figure, along with an estimated event size at that time of $\sim 1\,\mathrm{MB}$, represented a rough estimate of the rates that could be reasonably sustained through the offline processing stage. As the detector and software designs matured, CMS performed the first early estimates of the rate needed to carry out the main "discovery" physics program. The result is that a minimum of 80 Hz is needed (March 2002, LHCC presentations). When a few calibration samples are included, as was done for the more complete evaluation in the CMS-DAQ Technical Design Report [11], this figure became 105 Hz.

This figure of 105 Hz assumes that the experiment has reduced its rate to permanent storage to the bare minimum needed for it to maintain high efficiency for the well-studied Higgs, SUSY and Extra-dimension physics cases. Ongoing calibration and standard model studies (to be reported in the CMS Physics TDR) lead us to believe that an additional 50 Hz of trigger rate will permit CMS to complete the physics program, register adequate numbers of Standard Model events and guarantee that CMS can effectively look for any new physics offered by the LHC machine. This additional rate would be allocated mainly to jet channels; inclusive missing transverse energy; lowering of a few thresholds (e.g. the photon thresholds for the Higgs di-photon search so that more of the standard-model background can be measured directly in the data); as well as a number of topics in top physics

Experience gained from previous experiments at hadron colliders indicates that a lot will be learned with the first collisions at the LHC. Many of these estimates will be firmed up at that time. Given the uncertainties of the rate estimates from the combination of physics generators and the detector simulation, as well as the uncertainties of the machine and experimental backgrounds, we choose to use the figure of 150 Hz for the best estimate of the rate required for the physics program to proceed.

Certainly, CMS plans to record the maximum rate that its resources will accommodate, given that additional rate is simply additional physics reach. There is, a priori, no reason to limit the output of the experiment to any particular figure since the physics content is ever richer. The above figures are simply the result of today's estimates on the type of environment that the experiment will encounter as well as an attempt to limit the output to a figure that could be reasonably accommodated by the computing systems that are currently being planned.

2.5.3 Reconstructed (RECO) Data

RECO is the name of the data-tier which contains objects created by the event reconstruction program. It is derived from RAW data and should provide access to reconstructed physics objects for physics analysis in a convenient format. Event reconstruction is structured in several hierarchical steps:

1. Detector-specific processing: Starting from detector data unpacking and decoding, detector calibration constants are applied and cluster or hit objects are reconstructed.
2. Tracking: Hits in the silicon and muon detectors are used to reconstruct global tracks. Pattern recognition in the tracker is the most CPU-intensive task.
3. Vertexing: Reconstruction of primary and secondary vertex candidates.
4. Particle identification: Produces the objects most associated with physics analyses. Using a wide variety of sophisticated algorithms, standard physics object candidates are created (electrons, photons, muons, missing transverse energy

and jets; heavy-quarks, tau decay).

The normal completion of the reconstruction task will result in a full set of these reconstructed objects useable by CMS physicists in their analyses. They will be fully accessible by any members of the collaboration, who would only need to rerun these algorithms if their analysis needs required them to take account of such things as trial calibrations, novel algorithms etc.

Large scale event reconstruction will generally be performed by a central production team, rather than by individual users, in order to make effective use of resources and to provide samples with known provenance and in accordance with CMS priorities.

CMS production will make use of data provenance tools to record the detailed processing of production datasets and these tools will be useable (and used) by all members of the collaboration to allow them also this detailed provenance tracking.

Reconstruction is expensive in terms of CPU and is dominated by tracking. The RECO data-tier will provide compact information for analysis to avoid the necessity to access the RAW data for most analysis. Following the hierarchy of event reconstruction, RECO will contain objects from all stages of reconstruction. At the lowest level it will be reconstructed hits, clusters and segments. Based on these objects reconstructed tracks and vertices are stored. At the highest level reconstructed jets, muons, electrons, b-jets, etc. are stored. A direct reference from high-level objects to low-level objects will be possible, to avoid duplication of information. In addition the RECO format will preserve links to the RAW information.

The reconstructed event format (RECO) is about 250 KByte/event; it includes quantities required for typical analysis usage patterns such as: track re-finding, calorimeter reclustering, and jet energy calibration.

The access to all physics objects stored in the RECO format will be provided in a uniform way (interface) which will allow to retrieve the configuration (parameters) used for reconstruction.

This estimated RECO event size is consistent with the size of our current actual "DST" format. Only one RECO format will be supported but the ability to store multiple collections of objects reconstructed with different algorithms (versions) will be possible.

2.5.4 Analysis Object Data (AOD)

AOD are derived from the RECO information to provide data for physics analysis in a convenient, compact format. AOD data are useable directly by physics analyses. AOD data will be produced by the same, or subsequent, processing steps as produce the RECO data; and AOD data will be made easily available at multiple sites to CMS members. The AOD will contain enough information about the event to support all the typical usage patterns of a physics analysis. Thus, it will contain a copy of all the high-level physics

objects (such as muons, electrons, taus, etc.), plus a summary of the RECO information sufficient to support typical analysis actions such as track refitting with improved alignment or kinematic constraints, re-evaluation of energy and/or position of ECAL clusters based on analysis-specific corrections. The AOD, because of the limited size that will not allow it to contain all the hits, will typically not support the application of novel pattern recognition techniques, nor the application of new calibration constants, which would typically require the use of RECO or RAW information.

The AOD data format at low luminosity will be approximately 50 kB/event and contain physics objects: tracks with associated Hit's, calorimetric clusters with associated Hit's, vertices, jets and high-level physics objects (electrons, muons, Z boson candidates...). The AOD size is about 5 times smaller than the next larger (RECO) data format. Historically this factor is about the size reduction at each step that can both give important space and time improvements yet still yield sufficient functionality. This size estimate is consistent with our current prototyping of this event format. New versions of the AOD may be produced very often as the software and physics understanding develops.

Although the AOD format is expected to evolve in time, with information being added to assist in analysis tasks but also being reduced as the understanding of the detector is improved, this size is not expected to change significantly, especially when the potential use of compression algorithms is taken into account.

2.6 Event Data Flow

This section describes a baseline that allows the rest of the model to be sized in a coherent way. The HLT farm will write events at the maximum possible data rate that can be supported by the computing resources. Trigger thresholds may be adjusted up or down to match the maximum data rate, in order to maintain consistency with the downstream data storage and processing capabilities of the offline systems. All backlogs accumulated during periods of running at peak rates must be absorbed within 24 hours. We assume that in Heavy Ion running periods, CMS writes data from the online farm at the same rate (225 MB/s).

The proposed offline system will be able to keep up with a data rate from the online of about 225 MB/s. The integrated data volume that must be handled assumes 10^7 seconds of running. No dead-time can be tolerated due to the system transferring events from the online systems to the Tier-0 centre; the online-offline link must run at the same rate as the HLT acceptance rate.

The CMS baseline is that the online-offline link and Tier-0 are able to keep up in real-time with the HLT output rate. The actual LHC duty cycle (unknown, but possibly of order 50%) will thus result in some available contingency. This contingency is offset by factors that we reasonably expect to be present such as: link downtimes; event reconstruction

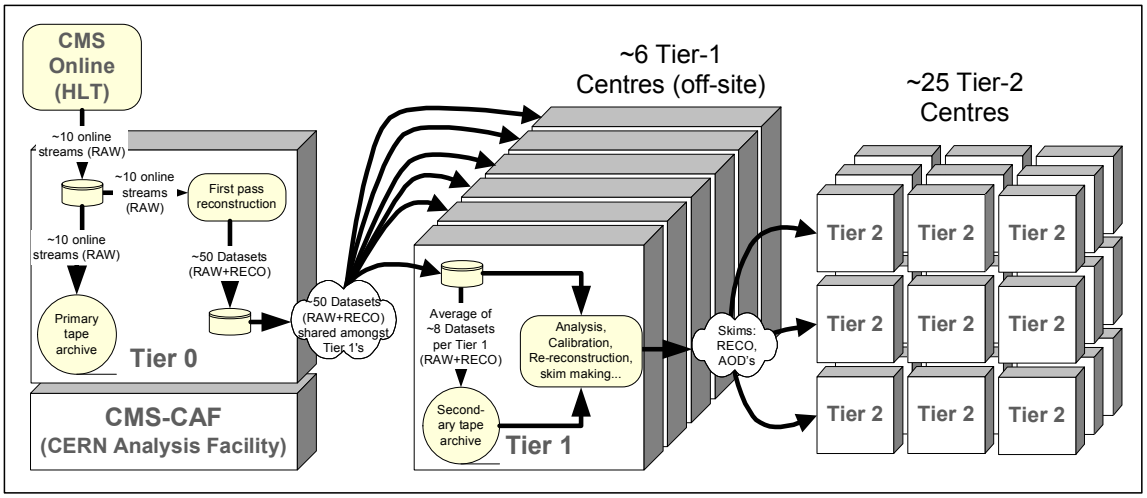


Figure 2.1: Schematic flow of bulk (real) event data in the CMS Computing Model. Not all connections are shown - for example flow of MC data from Tier-2's to Tier-1's or peer-to-peer connections between Tier-1's.

time uncertainties; actual event rates and sizes; Tier-0 downtimes; etc.

2.6.1 Data Streams

Figure 2.1 shows the Computing Centres in CMS Computing Model and the schematic flow of the real event data. The CMS online (or HLT) farm processes events from the DAQ system which have successfully passed the L1 trigger criteria. An entire event is distributed to an HLT node which either rejects it forever, or accepts it based on it passing one or more of the HLT selection criteria (the HLT trigger table).

The online system will temporarily store RAW events selected by the HLT, prior to their secure transfer to the offline Tier-0 centre. This raw event data constitutes the output of the HLT farm. To optimise data handling, raw events are written by the HLT farm into files of a few GB size.

The online system will classify RAW events into $\mathcal{O}(50)$ primary datasets based solely on the trigger path (L1+HLT); for consistency, the online HLT software will run to completion for every selected event. The first attribute of an event that is useful to determine whether it is useful for a given analysis is its trigger path. Analyses rarely make use of more than a well defined, and small, number of trigger paths. Thus events will be clustered into a number of primary datasets, as a function of their trigger history. Datasets greatly facilitate prioritisation of first-pass reconstruction, the scheduling of re-calibration and re-reconstruction passes, and the organisation of physics analysis.

For performance reasons, in the HLT Filter-Farm, we may choose to group sets of the $\mathcal{O}(50)$ primary datasets into $\mathcal{O}(10)$ online streams with roughly similar rates. The subdi-

vision of events into online streams can allow prioritised processing of a calibration stream which will result in updated calibration constants to be used for all subsequent processing for that data-taking period. Processing of certain lower-priority online streams may be deliberately delayed in the event of a partial disruption of service at the Tier-0.

The primary dataset classification shall be immutable and only rely on the L1+HLT criteria which are available during the online selection/rejection step. The reasoning for not rejecting events during re-processing is to allow all events to be consistently classified during later re-processing with improved algorithms, software, and calibrations.

The immutability of primary datasets in no way precludes the possibility of forming subsets of these primary datasets for some specific analysis purposes. For example it is expected that subsets of events that further satisfy some more complex offline selection can be made. These subsets may be genuine secondary event collections (formed by actually copying selected events from the primary datasets into new secondary datasets) or they may be in the form of event directories (lists of event numbers/pointers satisfying these selection conditions).

Duplication of events between primary datasets (mostly due to events also placed in an express-stream) will be supported, within reason, up to a maximum of 10%. The advantage of writing some events into multiple datasets is to reduce the number of datasets to be dealt with for a specific purpose later on (e.g. analysis or re-reconstruction). It facilitates prioritisation of reconstruction, application of re-calibration and re-reconstruction, even if distributed. The total storage requirements will not increase excessively as a result

The online system will write one or possibly several express-line stream(s) at a rate of a few % of the total event rate, containing (by definition) any events which require very high priority for the subsequent processing. As the name indicates the sole purpose of this stream is to make certain events available offline with high priority and low latency. The express-line is not intended for final physics analysis but rather to allow for very rapid feedback to the online running and for "hot" and rapidly changing offline analyses. Typical content of the express line could be: events with new physics signatures; generic anomalous event signatures such as high track multiplicities or very energetic jets; or events with anomalously low/high activity in certain detectors (to study dead/noisy channels). All events in the express-line are also written to a normal online stream and primary dataset, in order that they can be used in standard full analyses.

2.7 Heavy Ion Event Data

The CMS computing requirements are dominated by those required for pp physics; however CMS is also approved for running in heavy ion collisions and has a physics program targeted at this interesting area of study. Heavy ion runs are assumed to follow each major pp operation period as described in table 2.1.

The data rate (MB/s) for Heavy Ion running will be approximately the same as that of pp running although the event sizes will be substantially higher. To develop the computing requirements in this document we have taken a fairly conservative estimate for the average value of $dN/d\eta = 2000$. Event size estimates then using the same methods as for pp running, are estimated to be in the range 5-10 MB. We have also considered a mix of event types and event processing times per type that give a weighted mean processing time of about 10 times that for pp events. If more processing power were available, full reconstruction could be used and reconstruction times could be of order 5 times longer than this for a central event - and would yield in turn a richer physics program.

Due to the substantial reconstruction processing requirements for Heavy Ion events, and the very short running periods, it is not foreseen to follow their reconstruction in real time at the Tier-0. Rather a fraction of the events will be reconstructed in real time, while the remainder will be processed after the LHC run is complete. We aim to complete this reconstruction in a time similar to the LHC downtime between major running periods.

We currently consider three likely scenarios for this initial processing. (a) at the Tier-0 between pp running periods, (b) at one or more dedicated Tier-2 centres during a 4-6 month period between Heavy Ion runs, (c) some combination of (a) and (b).

The estimation of event sizes and processing times is not trivial. There are many unknowns, such as the mean multiplicity and the mix of events in the trigger. CMS expects to use regional reconstruction to keep the reconstruction time as low as possible. However we are following the latest results from RHIC that show that we could extend the physics reach of the CMS-HI running, by reconstructing more of each event.

2.8 Non-event data

CMS will have 4 kinds of non-event data: Construction data, Equipment management data, Configuration data and Conditions data.

Construction data includes all information about the sub detector construction up to the start of integration. It is available from the beginning of CMS and has to be available for the lifetime of the experiment. Part of the construction data is included also in other kinds of data (e.g., initial calibration in the configuration data).

Equipment management data includes detector geometry and location as well as information about electronic equipment. They need to be available at the CMS experiment for the online system.

Configuration data comprises the sub-detector specific information needed to configure the front-end electronics. They are also needed for reconstruction and re-reconstruction.

Conditions data are all the parameters describing run conditions and logging. They are produced by the detector front-end. Most of the conditions data stay at the experiment

and are not used for off-line reconstruction, but part of them need to be available for analysis. At the CMS experiment site there will be two database systems. The Online Master Data Storage (OMDS) database is directly connected to the detector and makes available configuration data to the detector and receives conditions data from the Detector Control System. The Offline Reconstruction Conditions DB ONline subset (ORCON) database is a replica of, and has information from the OMDS but synchronisation between the two is automatic only for conditions data coming from the detector. Configuration data is manually copied from ORCON to OMDS. ORCON is automatically replicated at the Tier-0 centre, to and from the Offline Reconstruction Conditions DB OFFline subset (ORCOFF) which is the master copy for the non-event data system. The relevant parts of ORCOFF that are needed for analysis, reconstruction and calibration activities are replicated at the various CMS computing centres using technologies such as those being discussed in the LCG3D project [12]. Details of the architecture, technologies, and designs are discussed in a roadmap document being written for CMS Calibration, Alignment and Databases. A CMS Note will follow with details of the specific implementation details that are chosen.

Estimates for the data volumes of the non-event data are being collected based on the anticipated use cases for each sub-system. This will be addressed in the first volume of CMS Physics TDR. Although the total data volume is small compared to event data, managing it carefully and delivering it effectively is essential.

Chapter 3

Computing System

In this chapter, we describe the computing centres available to CMS around the world, and the tiered architecture which will allow them to be used as a single coherent system. We specify the services that computing centres must provide, and give an overview of the performance requirements. The functionality of the CMS Tier-0 and Tier-1 centres has been described in detail in the CMS Computing Model [2], and is summarised here. We provide additional detail on the roles of the Tier-2 centres and the CMS CERN Analysis Facility (CMS-CAF).

3.1 Tiered Architecture

The CMS offline computing system is arranged in four tiers (see Fig. 2.1). The system is geographically distributed, consistent with the nature of the CMS collaboration itself. By following such an approach, CMS not only gains access to the valuable resources and expertise which exist at collaborating institutes, but also benefits from improvements in robustness and data security, through redundancy amongst multiple centres.

- A single **Tier-0 centre** at CERN accepts data from the CMS Online Data Acquisition System, archives the data and performs prompt first pass reconstruction.
- The Tier-0 distributes raw and processed data to a set of large **Tier-1 centres** in CMS collaborating countries. These centres provide services for data archiving, reconstruction, calibration, skimming and other data-intensive analysis tasks.
- A more numerous set of **Tier-2 centres**, smaller but with substantial CPU resources, provide capacity for analysis, calibration activities and Monte Carlo simulation. Tier-2 centres rely upon Tier-1s for access to large datasets and secure storage of the new data they produce.

- **Tier-3 centres** provide interactive resources for local groups and additional best-effort computing capacity for the collaboration.
- A **CMS-CAF** centre at CERN provides fast turnaround computing services local to the experiment.

The majority of CMS users will rely upon Tier-2 or Tier-3 resources as their base for analysis, with the Tier-1 centres providing the large-scale facilities necessary to support their work 'behind the scenes'.

3.2 Policies and Resource Management

The CMS computing system is composed of a large number of semi-autonomous centres, operating with resources volunteered from a diverse set of national, regional and institutional sources. Many of these centres may be shared or managed in common with other experiments or activities. In order to construct a working system, it is necessary to carefully define, in both a technical and managerial sense, the interactions between centres, and the services and resources they each provide. The resources contributed to CMS must be accounted for, and in many cases will be subject to formal Memoranda of Understanding.

It is a goal of the CMS computing system to allow maximum flexibility and freedom for users and analysis groups. However, it is inevitable that heavy resource contention will occur, especially during early LHC running. The system as a whole must therefore allow the specification of top-down policies governing data placement and priorities for use of resource. These policies are expected to change frequently as the LHC physics scenario becomes clear. Computer centres will work together to implement the CMS priorities on the fraction of the their systems declared to be 'CMS common resources'. Some of the tools and approaches which will allow this are specified in Chapter 4.

CMS plans to use Grid tools and infrastructure wherever it is feasible and appropriate. All 'common resources' must be accessible to CMS through agreed WLCG services. Countries and institutes contributing resources to CMS must implement these services such that any CMS user can potentially use their facilities.

3.2.1 Tier-0 Resources

The Tier-0 centre at CERN is by definition a CMS common facility, and is used purely for controlled production purposes.

Responsibility for setting up and operating the Tier-0 lies entirely with the CERN IT Division within the Worldwide LHC Computing Grid (WLCG) project [13]. The current

plan is that the facility will be a shared one, with the required minimum of resources for each LHC experiment guaranteed at all times. This will be implemented by an appropriately configured batch system. A detailed description of the proposed architecture of the Tier-0, in both hardware and software, is given in the WLCG TDR [14]. This also documents the choice of initial solutions, hardware life-cycle considerations, and the anticipated cost.

CMS will express its requirements for the Tier-0 within Service Level Agreements (SLAs) covering resource sizing and performance and the specification of the interfaces between the Tier-0 services and the CMS software which makes use of them. The CMS Computing Model contains more detail of operating scenarios, specifications and parameters. A WLCG Memorandum of Understanding [13] will define the responsibilities of CERN as host laboratory, including the Tier-0 implementation and operation.

3.2.2 Tier-1 Resources

CMS has identified a set of regional computing centres that will provide Tier-1 resources to CMS. At the time of writing, the centres with declared resource intentions, are: ASCC (Taipei), CCIN2P3 (Lyon), FNAL (Chicago), GridKA (Karlsruhe), INFN-CNAF (Bologna), PIC (Barcelona), and RAL (Oxford). There are also statements of intent from China, Korea and from the Nordic countries to host Tier-1 centres for CMS. The ongoing Tier-1 resource contributions and expected service levels are agreed in the context of the WLCG MoU [13].

Tier-1 centres provide both regional and global services. They fulfil many functions which are CMS common services, necessary for physics activities throughout the collaboration (e.g., data curation, reconstruction). Data will be placed at Tier-1 centres in accordance with explicit collaboration policy. In addition, a given Tier-1 centre may hold the only available copy of a given sample, and must potentially allow any CMS user to access it. These global services will be subject to the policies and priorities set by the CMS collaboration.

A given centre may also have responsibilities to users in its 'local community'; these functions must be managed so as not to have an impact upon its ability to fulfil its responsibilities towards the whole of CMS. We define 'local community' to mean those associated with the funding body that has provided the resources for the centre in question. This term is used in relation to both Tier-1 and Tier-2 centres.

3.2.3 Tier-2 Resources

The CMS Tier-2 centres are expected to vary widely in size, and in organisational and support principles. A 'typical' Tier-2 centre could be a large computer cluster purchased and operated by an Institute, with associated storage and network resources; nominal parameters for such a centre are specified later in this chapter. A Tier-2 is defined by the services it performs, rather than by its capacity. However, efficiency considerations require that a Tier-2 centre be a reasonable fraction of the nominal size. To CMS users, all Tier-2 centres must be capable of being accessed in a similar way. This implies that to the outside world, a Tier-2 centre appears to be a single monolithic entity, though this imposes no requirements upon its internal organisation.

Tier-2 resource contributions are declared in an MoU, and are subject to CMS accounting. Having Tier-2 status means that a facility is fully integrated into the CMS global computing environment. The community operating the centre may access data samples and support services from CMS, in return for supplying common services for the whole collaboration. Tier-2 centres are expected to provide a professional and well-coordinated operation, though with less tight constraints on availability and reliability than a Tier-1 centre. Each Tier-2 has a well defined point of contact with the CMS collaboration, such that any management issues may be channelled correctly.

Each Tier-2 is associated with a particular Tier-1 centre, which provides it with data access and storage services. A Tier-2 will though be able to obtain data from any Tier-1 centre. The regional support system is also based around the Tier-1 centres, and a Tier-2 can expect to receive direct support from its host Tier-1. Data transfer is a prime example of where support and coordination between Tier-2 and Tier-1 centres is required. Tier-2's also receive support from CMS in running any CMS specific code. To facilitate this, there will be a defined set of supported operating systems, with all officially released code guaranteed to work on compliant systems.

At most Tier-2 centres, some fraction of the resources will be allocated to common CMS tasks, with the remainder under the control of the local community. A typical model would be for a Tier-2 centre to host the work of one or more analysis groups. These groups will be assigned by CMS, in consultation and in agreement, with the local community where the objective will be to maximise the overlap of interests of the local community and the analysis groups. All of these groups of people will be able to decide within their respective groups how best to use the resources available to them.

Expanding on this, we foresee three types of use of Tier-2 resources:

- **Local community use.** Some fraction of the Tier-2 centre resources will be fully under the control of the local community.
- **CMS controlled use.** Tier-2 resources will also be used for organised activities allocated top-down by CMS. Examples are organised Monte Carlo production as well as CPU and disk space provision for defined analysis groups. Example uses

might be detector calibrations, creation of common physics group data samples, Heavy-Ion reconstruction or a variety of studies according to the priorities of the analysis group.

- **Opportunistic use** by any CMS member. All Tier-2 centres are accessible using the standard CMS distributed computing tools via WLCG services, and all CMS members have access to otherwise unused resources at all Tier-2s. This class of users will clearly have lower priority than the local community or resident analysis group. For opportunistic use to be a practical method of accessing CMS resources job turnaround times must be reasonable, and some storage available at all times. The method used to ensure reasonable turnaround times will be decided upon by individual Tier-2 centres.

The number and size of Tier-2 centres available to CMS is still evolving at the time of writing. For planning purposes, we assume that the equivalent of twenty-five centres of nominal size will be online by 2008. The number of institutes is likely to be larger than this. In some cases, however, a group of collaborating institutes may choose to aggregate their resources into a federated Tier-2 centre, in order to provide a single coherent resource of the appropriate scale.

CMS may also have access to computing resources at Tier-2 sites in the WLCG system which are not based at a CMS institute. These sites will offer spare CPU capacity, but are unlikely to be considered for long term data placement by CMS. Monte Carlo generation is therefore a natural workload for such sites.

3.2.4 Tier-3 Resources

Tier-3 sites are (often relatively small) computing installations that serve the needs of the local institution's user community, and provide services and resources to CMS in a mostly opportunistic way. The facilities at these sites do not form part of the negotiated baseline resource for CMS, but can make a potentially significant contribution to the experiment's needs on a best-effort basis.

Tier-3's are not covered by CMS MOUs, do not have direct managerial communication with CMS and are not guaranteed direct support from the CMS computing project. Nevertheless, Tier-3 sites are an important component of the analysis capability of CMS as they provide a location and resources for a given institute to perform its work with substantial freedom of action. Within the WLCG, Tier-3 sites are expected to participate in CMS computing by coordinating with specific Tier-2 centres, and they will provide valuable services such as supporting software development, final-stage interactive analysis, or Monte Carlo production. In particular, CMS expects to make opportunistic use of Tier-3 sites to provide additional CPU power for Monte Carlo data production.

3.2.5 The CMS-CERN Analysis Facility (CMS-CAF)

CMS requires computing services at CERN other than those provided by the Tier-0 centre. The CMS CERN Analysis Facility will provide a combination of services similar to those provided by typical Tier-1 and Tier-2 centres. The most important function of the CMS-CAF will be to enable short-turnaround and latency critical data processing, which needs to be carried out to ensure the stable and efficient operation of the CMS detector. The prompt access to recorded data which is possible at the CAF will be important for many of these activities. A further important function of the CAF will be to provide analysis services similar to those available at Tier-2 centres. Data reprocessing and Monte Carlo generation may also be carried out at the CMS-CAF if required.

The CMS-CAF will be hosted by the CERN IT Division. The selection of services it offers will be under CMS control. The centre will be accessible to all CMS users, who will have equal priority access to the facility in order to carry out analysis and data processing tasks. In addition, all users will have interactive access for code development purposes, will be able to remotely submit analysis jobs to run at the CMS-CAF, and will have an allocation of local storage space for processed data. However, it is intended that users with access to a Tier-2 centre will use this in preference to the CMS-CAF facilities. This allows the collaboration to make the most efficient use of the computing system and ensures that the CMS-CAF will have maximum flexibility for critical short-turnaround work.

CMS is well aware that CMS-CAF will appear a very attractive place to perform analysis, due to the ready access to new FEVT data. However, it is important to understand that the finite IO bandwidth to the CERN tape system will prevent arbitrary random access by users to the entire CMS dataset, and so explicit policy will need to be made by the collaboration on data caching and placement for the CMS-CAF, in a similar way to that at Tier-1 and Tier-2 centres (the CMS-CAF does of course retain the unique ability to rapidly access any event, in principle). To enable the entire computing system to function efficiently, CMS is therefore committed to ensuring that the facilities outside CERN at the Tier-1 and Tier-2 level receive full support and the most rapid possible access to data, such that they naturally become attractive to the analysis community.

3.2.6 Examples of Computing System Use

In this section, we present three hypothetical 'use cases' of the CMS computer system, intended to clarify the way in which the system is intended to function to meet the needs of the collaboration.

3.2.6.1 'Mainstream analysis'

CMS physicist 'A' is a member of the Higgs Physics Group in 2009, with a particular interest in the search for the SUSY Higgs in a particular decay channel. Since the collaboration has judged this area to be a high priority in the year 2009, the Group has been allocated generous analysis resources at several Tier-2 centres; indeed, A's own University provides a large Tier-2 centre for CMS. However, most of the other collaboration members at the University are interested in heavy ion physics, and so the local community has agreed with CMS to host analysis and specialised reconstruction activity for heavy ion data. A therefore uses a remote Tier-2 on a different continent for his analysis, and shares the resources mainly with other collaborators from that local community working on the same or related Higgs channels.

This mode of working has several advantages for A. He can share relevant data samples with his closest collaborators, and in weekly analysis meetings, the group reaches collective decisions on how to make best use of the resources dedicated to them, splitting their CPU capacity between analysis and fast Monte Carlo studies. From time to time, the group collectively requests a transfer of modest amounts of AOD or RECO data from their assigned Tier-1 centre to support their analysis. When the data in question is not present at the Tier-1 'local' to their analysis centre, it is transferred from a remote Tier-1 transparently, with no intervention by the group. The transfers may take several days to complete in this case, but at least the group is able to monitor the progress of the transfer as it progresses.

The group has been working on a new analysis strategy for the 2009 run, which is about to start. When they have refined their approach using fast Monte Carlo generated locally, they request the generation of a large sample of detailed Monte Carlo events. This work is scheduled centrally, and the jobs run at Tier-2 and Tier-3 centres around the world. The resulting events are stored at the local Tier-1 in case they are needed by other groups, and automatically copied to the host Tier-2 centre. The group combines their new signal Monte Carlo with standard background samples, and with specialised background samples used within the Higgs group. These samples are commonly used throughout the collaboration, and so are already cached at the local Tier-1.

Since the analysis approach is judged to be working well, the group decides with the agreement of the Higgs Group coordinator to invest a considerable portion of their allocated resources in testing their code on a large amount of real data from the 2008 run. The data in question is spread across three primary streams, none of which happens to

be held at the local Tier-1. This is no problem - the group specifies the code to be run, and the datasets which need to be accessed, and the jobs are scheduled at the correct Tier-1 sites. Since second-pass reconstruction from 2008 is still in progress, the group are worried that they may have to wait a long time for their jobs to complete. However, they are fortunate, and their work is given a high priority; some second pass reconstruction jobs are therefore displaced from the relevant Tier-1 centres back to the Tier-0 within a day, and their work can proceed.

If the run over real data proves successful, A will request that the group be allowed to skim events from the much higher statistics 2009 data in real time, as the first-pass AOD arrives at the Tier-1 centres. The group's Tier-2 centre will subscribe to this data sample, and run analysis straight away. He is looking forward to seeing his histograms fill as CMS runs!

3.2.6.2 'Calibration study'

CMS physicist 'B' comes from an institute which shares responsibility in 2008 for the monitoring and calibration of the ECAL detector. Much of her time is spent running Data Quality Management (DQM) jobs to check that the automated prompt calibration system at the CMS-CAF is operating correctly. It is important that this system works effectively, since the calibration results are fed back to the Tier-0 reconstruction farm within an hour of data being taken, and are used to reconstruct the majority of the data. In order to monitor the performance of the calibration procedure, B requires a sample of events taken from the reconstructed online data with minimum delay. This is ensured by the presence of a low-statistics but very high priority express stream, which contains sufficient events to perform DQM. For minimum latency, the DQM jobs are also run at the CMS-CAF.

B comes from a smaller institute, which does not have the resources to operate a Tier-2 centre for CMS. They host a Tier-3 centre, which is where B can test new DQM code, submit jobs to the CAF when necessary, and work interactively with the output ntuples. When the Tier-3 is not being used for interactive work, it runs a variety of CMS Monte Carlo generation jobs at low priority. Next year, the institute will join together with others in the same region to form a federated Tier-2 which can also offer secure storage of far more data for local use than is currently possible.

Due to an unforseen problem, the calibration system fails to supply correct constants to the prompt reconstruction for a section of the ECAL over a period of a few days. The problem is dealt with swiftly, but the affected events have already been distributed to Tier-1 centres, and the problem cannot be corrected by regenerating the AOD. Since ECAL calibration is critical for several high-priority channels, the decision is taken in the collaboration to perform a second reconstruction pass on the affected runs. The required reconstruction runs at the Tier-1 centres, and additional data is added to the calibration and conditions databases to invalidate the previously reconstructed data.

B's group decides to introduce a more robust calibration algorithm to avoid future problems. However, the algorithm will need testing thoroughly with a very large event sample before it can be rolled out in the CMS first-pass reconstruction. Moreover, the algorithm requires access to raw data for testing. The group submits a request for a large sequential pass over the previous month's data; this is granted by the collaboration, with moderate priority. The required jobs run at Tier-1 centres over the next few weeks; since the required access to raw data is known well in advance, scheduling of tape reads can be made with high efficiency and low impact upon other work at the Tier-1. When the algorithm is validated, it is further tested upon a dedicated online stream from the Tier-0 before being used for mainstream reconstruction.

3.2.6.3 'Hot channel'

Away from his day job, CMS physicist 'C' maintains an interest in particle physics theory outside the mainstream. A recent preprint suggesting that an unusual new GUT formulation may yield rare but observable charge-violating decays at LHC energies stimulates his enthusiasm; he decides to search for such events in the CMS dataset.

The coordinator of the relevant physics group has limited resources, and grants C's search only low priority. However, C has access to local resources at his institute's Tier-2, and can use these to carry out his study. The class of events he is looking for contain a single high-pt charged lepton, and all lie within a single primary dataset. He requests that the AOD for that dataset for the last three years be sent from his local Tier-1 to the Tier-2, and gradually it transfers.

When C carries out his analysis, he is amazed to find the exact signature he is looking for. He contacts his colleagues at other institutes, who replicate his results using their own code, submitting their jobs to his Tier-2 in order to gain access to his data sample. The group coordinator begins to take an interest, and grants priority for a more detailed study using RECO data to take place at the Tier-1 holding the relevant primary dataset; in fact, that Tier-1 centre is heavily loaded with reconstruction, and so the RECO is replicated to a second Tier-1 to enable the study to proceed quickly. The detailed analysis confirms the result, and word starts to spread throughout the collaboration.

C's work is given 'hot channel' status, and is granted high priority. The event class of interest is added to the express channel at the Tier-0, and the analysis jobs run over RECO data at the CMS-CAF within an hour of the data being recorded.

Alas, the inevitable soon comes to pass, and as C's result is scrutinised, it is identified as the result of a subtle efficiency bias in the CMS trigger system. This has the potential to cause major problems for the collaboration, since important results in other highly interesting channels are about to be presented at conferences, and could be affected. The collaboration takes the decision to immediately devote all resources to reprocessing of selected datasets relevant to the forthcoming presentations. The computing system

is able to rapidly react to this change of policy. This result is that the majority of events recorded by CMS are not reconstructed by the Tier-0 in real time, although the ongoing RAW data coming from the online systems are safely stored as usual at both Tier-0 and Tier-1. Spare capacity at the Tier-0 and all resources at Tier-1 are devoted to reconstruction, and the relevant physics groups given priority for analysis at Tier-2. After this effort results in the successful revision of the conference results, the Tier-0 resumes normal processing, though the backlog of stored events will require additional use of Tier-1 resources for reconstruction.

3.3 Tier-0 Centre

3.3.1 Tier-0 Functions

The Tier-0 centre is devoted entirely to the storage and sequential reconstruction of raw data. During CMS data-taking periods, the centre accepts and buffers data from the online systems at the CMS experimental site. Since no large-scale data storage facilities exist at the site, a safe temporary copy of the data is immediately made at the Tier-0 to allow the online system to release buffer space. After generation of corresponding RECO data by the Tier-0, the RAW and RECO components are stored together as FEVT data for simplicity of later access. The FEVT is copied to permanent and secure mass storage within both the Tier-0 and at an outside Tier-1 centre.

The CPU capacity of the Tier-0 is specified such that reconstruction keeps pace with the instantaneous data rate from CMS during p-p running at peak LHC luminosity. This allows spare capacity during running periods to recover from any backlog, as well as providing an overall contingency against currently unforeseen factors which might affect reconstruction times or data sizes (e.g. higher than predicted track multiplicities). We note that access to prompt calibration data is an important operational factor in achieving the necessary throughput.

CMS requires that the scheduling and completion of reconstruction tasks on the Tier-0 resources are independent of other CMS or non-CMS activities at CERN, and that therefore the Tier-0 will not contend for resources (e.g. CPU slots, disk buffers, tape IO) with any other activity.

For p-p data, CMS will use its Tier-0 resources to complete the initial reconstruction for all data taken in a given year. Heavy-ion reconstruction will be tailored to fit into the available time between the heavy-ion running period and the next p-p period - typically, about three months. Thus CMS anticipates essentially flat and continuous use of the Tier-0 farm for first pass reconstruction.

The final role of the Tier-0 is the distribution of FEVT data to external Tier-1 centres. 'First pass' AOD data will be produced as an adjunct to the main reconstruction, and this

is also distributed. Reliable distribution of FEVT for both data safety and analysis is of the highest priority for the experiment, and a robust handshaking mechanism is employed to ensure that all data is copied to at least two secure archives before being deleted from buffers at the Tier-0.

All aspects of Tier-0 operation will be subject to Quality of Service requirements. In general, 24/7 (around the clock, seven days per week) operation of the centre is required during data taking periods, and suitable fail-over architectures are expected to be used to ensure this.

3.3.2 Tier-0 Services

The Tier-0 does not directly provide 'user-visible' services to CMS collaborators. In order to support analysis, it must provide the following high-level services:

- **Acceptance of raw data:** The raw data from the experiment must be transferred from the online systems with guaranteed integrity and with low latency. The transfer system must have sufficient redundancy and capacity to handle backlogs that may occur during system operation. There are several technologies that can fulfill this data transfer task, e.g., CDR [15], PhEDEx [16]. The data transfer tools will interact with the Tier-0 monitoring as well as raise the appropriate alarms when operation exceptions occur.

- **Reconstruction of raw data:** The centre must perform in pseudo-real time the reconstruction of the raw data, stream the output into physics datasets and secure both RAW and RECO data onto tape as FEVT format. Monitoring of the reconstruction process must be possible in order to react quickly to faults or backlogs. It is the current assumption that reconstruction jobs will be steered using a batch system. This simple approach requires that latency trade-offs are well understood, e.g., the startup latency of a job versus the number of online stream files processed by it. Alternative means of operating the Tier-0 CPU farm, similar to those employed in the CMS online farm [11] could be adopted if required.

- **Mass storage:** The mass storage system which accepts FEVT data from the Tier-0 facility must provide guaranteed throughput for writing new files, and must provide monitoring and handshaking services. The mass storage system will need to be optimised for both essentially write-only use during data-taking periods, and heavy read-intensive use during offline periods.

- **Distribution of raw data:** The FEVT data must be reliably transferred to external Tier-1 centres, with guaranteed bandwith available to ensure no significant backlog or latency.

- **Prioritisation:** The Tier-0 must be capable of implementing CMS priorities on reconstruction and data distribution, and to accept new prioritisation decisions with short turnaround. An example of such prioritisation would be the processing and distribution of the 'express line' online stream with minimum delay in the event of a reconstruction or transfer backlog.

3.3.3 Tier-0 Resource Requirements

The Tier-0 performance requirements were estimated in the CMS Computing Model [2]. The 2008 requirements have been summarised here, and updated to reflect the current assumptions for the LHC running periods:

WAN: The transfer capacity for the Tier-0 centre is dominated by the prompt copying of FEVT data to external Tier-1 centres, and is estimated at 5 Gb/s

CPU: 4.6 MSI2K, sized according the expected data rate from CMS during p-p data-taking periods.

Disk: 0.4 Petabytes, required mainly for input buffer and distribution buffer space.

Mass storage: 4.9 Petabytes, sized according to the expected total sample size from CMS. Mass storage throughput during data-taking is estimated at 300 MB/s.

3.4 Tier-1 Centres

3.4.1 Tier-1 Functions

The CMS Tier-1 centres provide a wide range of high-throughput high-reliability computing services for the collaboration, based at large sites around the world. These centres provide their services through both WLCG-agreed Grid interfaces and higher-level CMS services, and are expected to supply very high levels of availability, reliability and technical support.

The tasks carried out at Tier-1 centres principally relate to organised sequential processing of data and extraction of datasets to be further analysed at Tier-2 centres. A given Tier-1 centre may have the only available copy of some data samples, and must therefore allow access by any CMS user in accordance with the policies and priorities set by the collaboration. Different users, or groups of users, may of course be assigned different priorities by the collaboration.

CMS physicists may perform event selection, skims, reprocessing, and other tasks on Tier-1 centre computers, processing data that has been previously placed at that site by CMS.

Most users will move the results of such processing to another computer centre (typically a Tier-2 centre), although we expect that some special types of analyses, such as those related to calibration work, or requiring very large statistics, may be entirely carried out at Tier-1 centres.

A Tier-1 centre may also choose to provide services to its local community, alongside the common Tier-1 services. These user groups will use the Tier-1 in the usual way, but will also have access to local long-term storage and local batch and interactive facilities for their analysis activities. This may be viewed as essentially the co-location of a Tier-2 centre at a Tier-1 site, with some sharing of management and infrastructure. This will be a useful mode of operation, but is in addition to the common CMS Tier-1 functionality, and will not be permitted to interfere with the role of the centre as a common facility with workload under CMS prioritisation. We note again that we expect that most CMS users will use a Tier-2 or local Tier-3 centre as their host facility for CMS analysis.

We note that in many cases Tier-1 centres available to CMS will be shared between experiments, and that many WLCG services at the centre will be common between activities. However, each Tier-1 centre must provide CMS specific services where required.

A second crucial function of Tier-1 centres is to provide a large portion of the experiment raw and simulated event data storage. CMS intends to make two 'custodial' copies of all raw event data. The first will be stored at the CERN Tier-0 centre, and the second distributed between Tier-1 centres. Neither of these copies will be regarded as purely 'backup', and both will be used for reprocessing when required in order to optimise the overall efficiency of the system. The acceptance of a portion of the raw data by a Tier-1 site implies that the centre undertakes its responsible stewardship in the long term. This involves ensuring that the underlying mass storage system is protected against controllable risks such as hardware or media failures, environmental hazards, or breaches of security.

In addition to storing CMS data, the Tier-1 centres take responsibility for distributing data to analysis applications as required, and must provide suitable high-performance IO and network infrastructure. The Tier-1 centres are expected to serve FEVT data to Tier-2 centres for analysis, and Tier-1 centres for replication, upon demand, with prioritisation set by CMS. Tier-1 centres are also expected to accept data from Tier-2s. This includes simulated data samples generated at Tier-2 and Tier-3 sites, and derived data produced during Tier-2 analysis that is needed at other sites, or that would be inefficient to reproduce in the case of Tier-2 storage failure. Connections between Tier-0 and Tier-1, and between Tier-1s, are expected form the 'backbone' of the CMS data distribution system.

3.4.2 Tier-1 Services

Tier-1s will provide the following set of user-visible services:

- **Data Archiving Service:** The Tier-1 centres are expected to archive a share, commensurate with the size of each centre, of a copy of the raw data, and the primary copy of the simulated data. The Tier-1 centre takes responsibility for active maintenance and security of these data samples. The availability of these data will be the subject of service level agreements with the collaboration.

- **Disk Storage Services:** The Tier-1 centre will provide large amounts of fast disk storage to act as a cache to mass storage archives, as a buffer for data transfer, and to provide rapid access to the entire AOD sample for analysis and event selection.

- **Data Access Services:** The Tier-1 centres are expected to provide access to the data entrusted to them for analyses both on local CPU resources and at other computer centres. Flexible prioritisation of access to data and provision of accounting information is expected.

- **Reconstruction Services:** The Tier-1 centres are expected to provide resources for running second-pass reconstruction or other large-scale workflows with high-throughput requirements. It must be possible for CMS to prioritise these operation against other processing tasks.

- **Analysis Services:** The Tier-1 centres are expected to provide capacity for both direct analysis, and for processing in support of analysis elsewhere. In the latter case, this will primarily be the skimming and reduction of data samples for transfer to Tier-2 centres for further analysis.

- **User Services:** The Tier-1 centres may also provide Tier-2 type user services, alongside the common services.

The availability and quality of the common services will be the subject of formal service level agreements with the collaboration.

In order to meet the above requirements a Tier-1 centre must provide some specialised system-level services:

- **Mass storage system:** The implementation of the MSS systems will vary between sites; in all cases, the archive should be reliable enough to accept custodial responsibility for a copy of the raw data. The risk for data loss should be calculated and understood. In many cases, the system should be cached to achieve the required performance for data access.

- **Site security:** The site must provide infrastructure to enable security and provide appropriate access restrictions for resource use. Tier-1 centres are expected to take a coherent and proactive approach to security issues.

- **Prioritisation and accounting:** Access to data and resources must be arbitrated between users and user groups in accordance with CMS priorities and policies.

Changes of policy must be implemented with rapid turnaround. Accounting information must be recorded and provided to CMS for both resource usage and access to data samples.

- **Database Services:** Tier-1 centres must provide facilities for hosting large databases holding a variety of non-event data. The databases will be populated both via replication and caching of database information held centrally or at other sites, and as a result of operations at 'local' Tier-2s.

3.4.3 Tier-1 Resource Requirements

The estimated capacity of a nominal Tier-1 centre is given below. CMS Tier-1 centres will be of a range of sizes, with the average size likely to be below that of a nominal Tier-1 in 2008. In order for a Tier-1 centre to operate efficiently and to offer access to all AOD and an appropriate fraction of FEVT data, there is a natural lower threshold on capacity compared to the nominal size. The detailed calculation of the Tier-1 parameters is given in the CMS Computing Model [2], which also specifies the relevant efficiency factors. The calculation has been updated to reflect the assumptions given in Chapter 2. The evolution of nominal Tier-1 capacity with time is illustrated in Chapter 5.

WAN: The incoming transfer capacity for a nominal Tier-1 centre is 7.2 Gb/s. The incoming data is a combination of raw data transfers from the Tier-0 for custodial data storage, updates of reprocessed RECO and AOD samples from other Tier-1 centres, and transfers of processed and simulated data from the Tier-2 centres. The outgoing transfer capacity for a nominal Tier-1 centre is 3.5 Gb/s, in order to serve other centres requesting data hosted on Tier-1 storage resources. These capacities are in part driven by considerations of transfer latency as well as throughput. These estimates represent a minimum requirement; they imply a highly structured and controlled transfer environment. Much more benefit could be reaped from the distributed environment, along with much more stability against operational saturation, if these network capacities could be significantly increased, and indeed some Tier-1 centres are already planning to implement 2-4 times this capacity. CMS strongly endorses efforts to increase the available bandwidth.

CPU: The total processing capacity at a nominal Tier-1 centre is 2.5MSI2k. The processing is split roughly in the ratio of 2:1 between scheduled data reprocessing, and analysis activities.

Disk: The disk capacity of a nominal Tier-1 centre including efficiency factors is 1.2 PB. About 85% of the disk space is utilised for serving data for analysis. The remainder is used for reprocessing staging space and space for serving and staging simulated data.

Mass Storage: The estimate for mass storage (tape) at a Tier-1 centre is 2.8 PB. The data loss rate should be at least comparable to current tape based systems, that typically lose data at rate of 10s of GB per PB stored.

Data Rate from storage: The data access rate for the mass storage system is estimated at 800 MB/s. Most data in mass storage is written once and read many time

CPU node I/O bandwidth: Gigabit connectivity is required (See discussion also for Tier-2 below)

3.5 Tier-2 Centres

3.5.1 Tier-2 Functions

Each Tier-2 centre provides CMS with a flexible resource with large processing power, but with looser storage, availability and connectivity requirements than a Tier-1. This allows such a centre to be provided at reasonable cost by an institute or group of institutes.

The basic functions supported by a Tier-2 include:

- Fast and detailed Monte Carlo event generation.
- Data processing for physics analyses, including late stage analysis requiring very fast data access.
- Data processing for calibration and alignment tasks, and detector studies.

To users, a Tier-2 appears as a single entity. The community that runs the Tier-2 may organise it any way they see fit, as long as this requirement is met. It is anticipated that most Tier-2s will consist of a set of computing resources located at a single site, with very good (LAN quality) connectivity amongst the components, and the resources managed by a single support team. There is nothing to prevent a Tier-2 from consisting of computing resources that are physically remote and hence connected using a WAN, as long as the Tier-2 integrates them efficiently and still defines a single point of management contact with CMS.

For example, a Tier-2 may choose to perform Monte Carlo simulation tasks at a remote auxiliary facility that is available only 50% of the time. The Tier-2 community would be expected to manage the associated complexity and ensure that this mode of operation appears transparent to CMS. In this spirit, we anticipate that smaller institutes will contribute to Tier-2 computing by either pooling together to make a viable 'federated' Tier-2 or by associating with an already existing Tier-2. It is up to each local community to determine the most efficient operational strategy.

3.5.2 Tier-2 Services

User-visible services required at each Tier-2 centre include:

- Medium- or long-term storage of required data samples. For analysis work, these will be mostly AOD, with some fraction of RECO. RAW data may be required for calibration and detector studies.
- Transfer, buffering and short-term caching of relevant samples from Tier-1's, and transfer of produced data to Tier-1's for storage.
- Provision and management of temporary local working space for the results of analysis.
- Support for remote batch job submission.
- Support for interactive bug finding e.g. fault finding for crashing jobs.
- Optimised access to CMS central database servers, possibly via replicas or proxies, for obtaining conditions and calibration data.
- Mechanisms for prioritisation of resource access between competing remote and local users, in accordance with both CMS and local policies.

To support the above user-level services, Tier-2s must provide the following system-level services:

- Accessibility via the workload management services described in Section 4.8 and access to the data management services described in Section 4.4.
- Quotas, queuing and prioritisation mechanisms for CPU, storage and data transfer resources, for groups and individual users.
- Provision of the required software installation to replicate the CMS 'offline environment' for running jobs.
- Provision of software, servers and local databases required for the operation of the CMS workload and data management services.

Additional services may include:

- Job and task tracking including provenance bookkeeping for groups and individual users.
- Group and personal CVS and file catalogues.
- Support for local batch job submission.

3.5.3 Tier-2 Resource Requirements

In order to give an example of the size and hardware requirements for a CMS Tier-2 centres, some useful numbers from the CMS Computing Model document and some derived ones are given here. These numbers apply to the reference year 2008. These numbers refer to a 'nominal' Tier-2, so are typical and average, though in practice we expect there to be a wide range of sizes of Tier-2 centre. In addition, the goals and internal infrastructure of Tier-2s will vary considerably; for instance, an individual Tier-2 may be more geared to analysis then Monte Carlo or vice versa.

CPU: 0.9 MSI2K.

Disk: 200 TB.

WAN At least 1 Gb/s. If a Tier-2 is structured as a collection of sparse centres, each of them, declared to be available for analysis activities, needs such connectivity. For Monte Carlo production the requirements are less stringent. It will be beneficial if at least some Tier-2 centres have 10 Gb/s connectivity, as the total processing power of a Tier-2 in 2008 (900 kSI2k) roughly implies such a transfer capability for an event size of 400 kB (the RecSimSize from the Computing Model). As with the requirements quoted above for Tier-1 WAN capacity, these estimates represent a minimum requirement. CMS strongly endorses efforts to increase the available bandwidth between Tier-2 centres and between Tier-1 and Tier-2 centres.

Data import: 5 TB/day of data import from Tier-1 for AOD and other data replicated at T2. This number gives an idea of the scope of the data management a T2 has to perform locally. It translates to a few thousand files to store every day, possibly replacing older copies. It implies that it takes roughly 20 days to refresh all the replicated data on disk at a Tier-2 centre in 2008. It will be beneficial if at least some of the Tier-2 centres support peak refresh rates an order of magnitude larger than this.

Data export: 1 TB/day. As above, this is to give a sense of data management needs. This requirement is derived from the Computing Model indication for $\sim 10^8$ simulated events per year per Tier-2, multiplied by the event size and divided by 200 days to obtain about 10^6 MB/day $\sim$ 1TB/day. This is an average; occasionally data may be produced faster and may need some local buffering. An upper limit is obtained if we assume a Tier-2 is used 100% for Monte Carlo generation (likely to happen for some limited time period) and we assume fast simulation. Each event simply takes the reconstruction time (25 kSI2K.sec/evt from the Computing Model), and the Tier-2 would then produce about 40evt/sec or $4 \times 10^6 \times 2$ MB/day $\sim$ 8TB/day

CPU node I/O bandwidth: 1 Gb/s. From 0.25 KSI2s/evt and 50 kB/evt (estimated analysis CPU and AOD size in the Computing Model) and a $\times 4$ safety factor, we derive 7 Mb/s/kSI2K. In practice we can expect the size of analysis input to be

anywhere between the estimated AOD number and the more conservative Reco size (0.25 MB/evt), bringing the input need to 32 Mbit/s/kSI2K. In 2007, one might expect compute nodes to have 8kSI2K thus suggesting gigabit connection from all worker nodes. In practice, it is reasonable to imagine a Tier-2 being built with gigabit connection to all the worker nodes. This will also allow Tier-2s to support fast, interactive style, analysis applications on data samples too large to fit on desktops.

Aggregated bandwidth: 10 Gb/s. Total requested bandwidth from disk to CPU for analysis. To be on the safe side a Tier-2 may want to guarantee that all its nodes can run the fast analysis application indicated above (still by no means the fastest task that can be imagined). This leads to an integrated disk-to-cpu bandwidth of order of 10Gb/s. It may be useful for at least a few Tier-2 centres to be capable of significantly exceeding this.

Jobs submission frequency: The average rate at which a Tier-2 must accept new jobs for execution. The rate for a single Tier-2 is not particularly demanding, other than in the case of massive job failure. This applies even for a relatively short job lifetime of six hours (the current value from CDF analysis farm which runs users' analysis and simulation but no organised production), and up to a thousand execution slots. The rate under these conditions is one new job every twenty seconds. Of course a global scheduler managing the integrated CMS analysis effort, will be required to manage about 50 to 100 times this frequency. Job submission from the user side is expected to come in bursts, and some action will be required to smooth out the load to the local batch manager.

3.6 CMS-CAF

3.6.1 CMS-CAF functions

Latency critical activities will uniquely be performed at the CAF. These are related to the efficient performance of the CMS detector. A particular advantage of the CMS-CAF in this respect is the access to the full RAW data sample which is stored at the CERN Tier-0 facility. These types of activity are likely to include:

- **Diagnostics of detector problems.** This service will be particularly important during early running and after shutdowns.
- **Trigger performance services** such as reconfiguration, optimisation and the testing of new algorithms. Such activities will be performed in response to a variety of circumstances such as changing or unexpected backgrounds, performance of a particular sub-trigger or the need to rapidly focus the trigger selection in on interesting physics.

- Derivation of **calibration and alignment data** with very short turnaround, for example to support the high level trigger algorithms and initial reconstruction at the Tier-0.

These activities will have the highest priority at the CAF and will take priority over all other activities.

In addition the CMS-CAF will provide a central service for the following:

- Central databases to support data management and other CMS-specific functions.
- Production bookkeeping and workflow records.
- Collaboration software and document repositories.
- Collaboration WEB services and Data-Publishing services

3.6.2 CMS-CAF Services

Services provided by the CMS-CAF will be a union of those offered by Tier-1 and Tier-2 centres outside CERN, detailed in the sections above. The CAF will support very rapid access to a subset of FEVT data similar in size to that at a nominal Tier-1, and access to the entire AOD sample. In addition, it will provide reconstruction services and interactive and batch analysis facilities in the same way as Tier-1 and Tier-2 centres respectively. The analysis facilities offered by the CMS-CAF will be approximately equivalent to two nominal Tier-2 centres.

In order to support its unique function, the CMS-CAF will require the following special services:

- Interactive login facilities capable of supporting the CMS policy of access by all collaborators.
- High-quality disk space provision for all CMS collaborators.
- A highly flexible batch queueing system capable of rapid implementation of new workload management priorities.
- An appropriately sized user support team providing services for a potentially large number of collaborators who need to work at the CMS-CAF in addition to a Tier-2 centre.

3.7 Current Status of Computing System

In this section, we give brief details of the status of the CMS computing system at the time of writing. The future development and deployment plan for the computing system is specified in Chapter 5. The key tests of the current CMS computing infrastructure have been in a series of yearly 'data challenges', culminating in 'DC04' which processed around 70 M events at Tier-1 and Tier-2 centres worldwide. The DC04 computing resources reached 25% of 2007 scale. Results of this and other CMS computing challenges may be found in the references given in Appendix B.

3.7.1 Tier-0 Centre

The detailed implementation of the Tier-0 centre at CERN is under study. The provision of such a centre is the responsibility of the CERN IT Division. Many of the technologies to be employed in the Tier-0 have been prototyped over recent years and are under development in order to reach the required levels of performance and reliability. Examples are CASTOR, QUATTOR and CDR [15]. Details of the Tier-0 deployment and prototyping are specified in the WLCG TDR [14].

3.7.2 Tier-1 Centres

At each of the six confirmed Tier-1 sites for CMS, technical development is taking place in order to establish the buildup of resources toward CMS running. Most of these sites already host substantial resources which are in frequent use for CMS Monte Carlo production and analysis. Over 100 M detailed simulation events have been produced since mid-2003, the majority at Tier-1 sites. CMS production and analysis are now moving to systems based upon WLCG services, as these services become well established at Tier-1 sites; details are given in Chapter 4.

All current CMS Tier-1 centres are integrated into the CMS data management system, and host data samples for local or remote analysis; in most cases, these are many TB in size. The majority of data has so far been transferred from CERN to Tier-1 sites in the current system, with sustained rates of over 700 MB/s having been achieved. Direct transfer of data between Tier-1 sites when required has now also begun on a regular basis. Serving of data by Tier-1 sites to Tier-2 sites is being tested at increasing rates, so far reaching the 10 MB/s level.

The main route for further development of operational Tier-1 WLCG services is through the WLCG Service Challenges, as detailed in the relevant TDR [14]. In parallel, the CMS-specific services, support systems, and management structure are being defined.

Tier-1 centres will form the backbone of the upcoming CMS computing challenges detailed in Chapter 5. All Tier-1 sites will be expected to participate in such challenges with the full range of required services.

3.7.3 Tier-2 Centres

The number of identified CMS Tier-2 sites has increased rapidly over the last year, though the final number is not yet confirmed. Many Tier-2 sites have already contributed to ongoing CMS Monte Carlo production and analysis, including sites not associated with a CMS institute, but offering capacity through the WCLG workload management system.

The upcoming WLCG Service Challenges will incorporate CMS Tier-2 sites, and will be the next step in permanent integration of such sites through WLCG services. In the ramp up to full LHC capacity in 2007/8, the role of Tier-2 sites will become steadily more important as the expected environment for user analysis.

3.7.4 CMS-CAF

The CMS-CAF combines the services of Tier-1 and Tier-2 centres with certain specialised services required to support a large user community and offer collaboration-wide functionality. CMS has provided this basic range of services for the collaboration for many years, through both the CERN public login service and dedicated CMS facilities. In particular, CERN is currently the main base for CMS analysis, and hosts the collaboration bookkeeping and workload management services. Most code development and testing is also performed at CERN.

The clear goal of CMS is to provide such services at a wide range of Tier-1 or Tier-2 sites, and to reduce the dependence of the collaboration upon special services provided from CERN. Nevertheless, the experience gained in hosting the current services forms a solid foundation for the development of the CMS-CAF. It is expected that the current CERN public login service, alongside enhanced CMS-specific services, will form the first step towards full CMS-CAF functionality towards the end of this year.

3.8 Deployment of Computing Services

This section summarises the set of services needed at each site and the computing infrastructure they require, focussing on CMS specific services beyond those we expect to be provided by WLCG. A similar set of services is required at each Tier-1, -2 and -3 site, although simpler and/or partial implementations may be sufficient at the smaller sites.

In this document we address the general functionality, and do not attempt to provide detailed quantitative requirements. We expect to provide realistic quantitative requirements on the time scale of WLCG Service Challenge 4 (SC4).

3.8.1 Computing Services Overview

A number of services are expected to be made available for CMS use at each participating site. These can be generally classified as User Interface (UI), Worker Node (WN) of the Computing Element (CE), gateway systems, and infrastructure services supporting these, such as databases and storage space.

Unless otherwise stated, CMS assumes that services will be provided around the clock. This include production services at all Tier-1 and -2 sites, with implied fail-over, mirroring, back-up and monitoring. Database services are expected to be isolated from those of other experiments or projects, such that overloading or failure of servers for other projects should not affect the quality of service for CMS.

The required services include:

- Storage space for hosting CMS software. The software storage must be accessible to all worker nodes, user interface systems and gateway servers. The storage area must be writable by CMS software administrators. No access from outside the site is required.

- Small database services. These are not for the event or condition data but databases for job monitoring, transfer system, etc. These databases are local to the site and do not require cross-site replication or inbound access.

- Local file catalogue. The catalogue typically requires a database as well. Read / write / update access is required from all worker nodes and gateway systems. No access from outside the site is required.

- Conditions data requires both storage capacity for the data itself and may require servers for running caches.

- Each site is expected to deploy one or more User Interface systems with public login access for physicists permitted to use the site. The user interface machines are for developing and testing code, and for submitting jobs.

- Worker nodes. Jobs on worker nodes require site-local configuration information such as where the file catalogue is and how to access condition data. However, the CMS software assumes as little as possible regarding the worker node environment, and does not require other services to run on worker nodes.

- CMS will install CMS-specific services on so called gateway systems, which are typically installed as UIs but provide login access only for a restricted number of administrative users. The gateway servers must have sufficient local disk space for installing the executables for services, and for storing small amounts of internal state and log information to decouple services from failures in offsite network access. The gateway services do not require substantial amounts of CPU capacity or network bandwidth. The gateway servers must have external outbound connectivity, access to local resources (storage, file catalogue, databases) and inbound connectivity from local worker nodes. CMS reserves the possibility to also require inbound external connectivity to specific gateway services; the collaboration will in this case negotiate reasonable security guarantees on such services.

3.8.2 User Interface

The UI is used for small-scale development, testing and debugging of user analysis applications, and job submission. This section summarises the requirements for these applications on the UI.

In general it is not required to have any servers running on the UI. The services provided should be available to a generic user, without any special privileges on the local system (e.g. adminstrative access):

- Software installed and configured for execution and development. There should be a mechanism to keep the software up to date with respect to a central repository, with automatic and/or on–demand installation and synchronisation. Eventually, it should be possible to select only a fraction of the whole software tree if only a subset of the full functionality is required (e.g.only 'analysis' software without simulation and/or reconstruction).

- It should be possible to access a small but representative data sample in order to test the analysis or reconstruction application interactively during the development phase. For environments such as a laptop, where disconnected operation is required, a straightforward way to download small, pre-defined samples should be provided for this purpose, perhaps sharing the same mechanism used for software installation.

- Access to CMS central databases. CMS will have several DB's that need to be accessed by every User Interface: for instance, detector condition, calibration, dataset bookkeeping, and dataset location. This access is likely to require a local cache server on the gateway. The UI must also provide a local disk area where a small fraction of the DB's can be replicated for local usage while the system is offline (including writes for later uploading of changes to main DB).

- CMS will have a service implementing job bookkeeping, using for instance multiple BOSS databases. The UI will be able to access them and optionally host such a

service for the local users. In this case, while a full database solution (e.g., MySQL) could be used, the database will usually be lightweight (e.g.SQLlite) and typically file based, so no DB server will be required.

- Access to WMS, including all necessary tools if the UI is required to submit to more than one variety of underlying Grid system.

3.8.3 Worker Node

On the worker node, the following services are needed, and should be located through a site-specific configuration mechanism:

- Local File Catalogue
- Local software repository, including the setup of the user environment as needed: this should be identical to that available on interactive machines.
- CMS DB servers, or a route to access them via cache.
- Definition of the 'close' SE, for staging of job output before transfer to the final destination.
- A mechanism for notifying relevant CMS agents of the presence of a new output file in Local File Catalogue to be inserted in DBS and moved to different storage as required.

3.8.4 Gateway Servers

The services CMS currently expects to host on the gateway servers are listed below. It is not excluded that more or fewer services will be included in future.

- Agents for monitoring integrity of data and other site information.
- Data placement and transfer agents, including monitoring agents for the 'CMS dashboard'.
- Data location service agents.
- Job monitoring and logging services, including optional real-time monitoring and relaying logging information out of the site (BOSS).
- Software installation management.
- Job output collation.

Chapter 4

CMS Computing Services And System Operations

4.1 Introduction

The CMS computing environment is a distributed system of computing services and resources that interact with each other as Grid services. The set of services and their behaviour together provide the CMS computing system as part of the Worldwide LHC Computing Grid. Together they comprise the computing, storage and connectivity resources that CMS uses to do data processing, data archiving, event generation, and all kinds of computing related activities.

Given the time scale and schedule for this Design Report, we will not attempt a complete engineering blueprint for the system and its components. Instead, we outline the principles of the architecture and approaches and in many cases give just a sketch that describes how the basic use cases and workflows are implemented in terms of system components and computing services. Where possible, quantitative metrics for performance and scales will be given, or in some cases instead indications for a plan how to eventually arrive at those numbers.

The baseline as described in this document will be the primary focus of development and will provide the basic functionality needed to achieve the goals of experiment at LHC turn-on. As of the time of the writing of this document, some of the components/services described already exist, some of them exist, but require evolution to play the roles described, and some of them need to be developed from scratch. In order to give a better sense of the status of the computing project, we will indicate the a brief summary of the status of each of the components at the time this document was written.

In some cases, we also give possible extensions beyond the baseline capabilities and functionalities. These "beyond the baseline" extensions may or may not be pursued after the

baseline system is deployed. We nonetheless include rough descriptions of these possibilities primarily to seed further post-CTDR discussions.

We realise that this way we do not give a complete specification of the system, but instead attempt to be as specific as possible to give the recipe and roadmap to arrive at a sufficiently functional and scalable system to support initial data taking. This document provides a snapshot of our understanding of the CMS computing services, which will be continuously evolved further after the release of this document, as a living document.

The general approach of CMS to developing our computing system is an iterative process of developing the system components, and integrating them together at successive steps of scale, testing these steps in major "challenges" (service challenges, magnet test, CSA, readiness for data taking), before taking the next step in adding functionality and scale of the system. As the lower-level Grid services evolve, the application level services will be adapted to take advantage, and new versions of the system will be integrated and released for production use.

To enable this procedure we adopt a loosely coupled system of services that can be improved upon and replaced with better versions (higher performance, more functionality), while specifying well defined interfaces and delegating functionality across the software stack. This approach allows us to commission increasingly functional and scalable systems in the absence of a fully-defined engineering blueprint of all the components.

The iterative approach will enable CMS to maintain a production-quality computing environment that gets continuously upgraded as required during the lifetime of the experiment.

4.2 General principles

In putting together the architecture of the baseline CMS computing system a number of guidelines were followed:

Optimisation for read access. In general event data in HEP is written once, never modified and subsequently read many times. CMS is unlikely to be any different. If the access patterns are roughly understood it is clear that paying some extra cost during the write step to ease subsequent reads makes sense.

Optimise for the large bulk case, but without limiting a basic user from accomplishing basic tasks. The largest computing problems that CMS will face come from the management of very large amounts of data and jobs in a large distributed computing system. The system *must* be optimised to support this bulk case well and appropriate infrastructure to support this must be developed and deployed. However it should be possible for users doing smaller studies in more restricted environments to obtain and access data. A user doing development/testing on a desktop is the canonical example, but also a user working

"locally" at a university site typically falls into this category.

Minimise the dependencies of the jobs on the Worker Node (WN). The most difficult scaling and service reliability problems usually appear in the environment of the WN. This is simply because we expect jobs to run 24 hours a day, 7 days a week on 10^3-10^4 worker nodes scattered around the distributed computing system. While outages or interruptions of services needed only for administration of the system (or even for creation and/or submission of new jobs) can be tolerated, we expect that at any given instant in time there will always be many jobs queued and jobs starting on WN's. By minimising dependencies, the overall throughput of the system can be made more stable and fault-tolerant. Similar considerations also lead one to the asynchronous handling of job output (relative to the job finishing). In addition any dependencies which do exist for an application on a WN should be local to the site to avoid single points of failure for the entire system.

Allow for provenance tracking. As a requirement on the software framework and the computing infrastructure it must be possible to track the provenance of datasets produced. It is likely that this is accomplished via a combination of cvs tags for the software and run-time parameter set(s) for the application, and input as well as output dataset specification for the workflow management system. The desired transformation from one dataset to another is thus unambiguous, and in principle reproducible.

Site-local configuration information should remain site-local. This adds flexibility for the local site system administrator to configure and evolve the local system as needed without synchronisation to the rest of the world. It requires, however, that some (hopefully simple and dependable) means be available for jobs to discover the site-local configurations when they start on a worker node.

Keep the solution simple and avoid paying the cost of complexity unless actually needed. This principle applies in a number of areas. In particular we note that it may result in offering users a set of options which allows them to trade reduced functionality for simplicity.

4.3 System overview

As noted above we envision that the CMS computing system will consist of a distributed set of systems and services. Many of these services will be provided by sites and present standard Grid interfaces as developed by Grid projects. Some of the services implementing particular CMS application behaviour will be VO-specific and are being developed within CMS, often in collaboration with Grid projects and based on Grid standards and interfaces. This includes CMS application services running at sites, specific experiment data services that know about CMS data structures and metadata, and services that implement CMS views of resources and execute CMS policies and behaviours, again typically

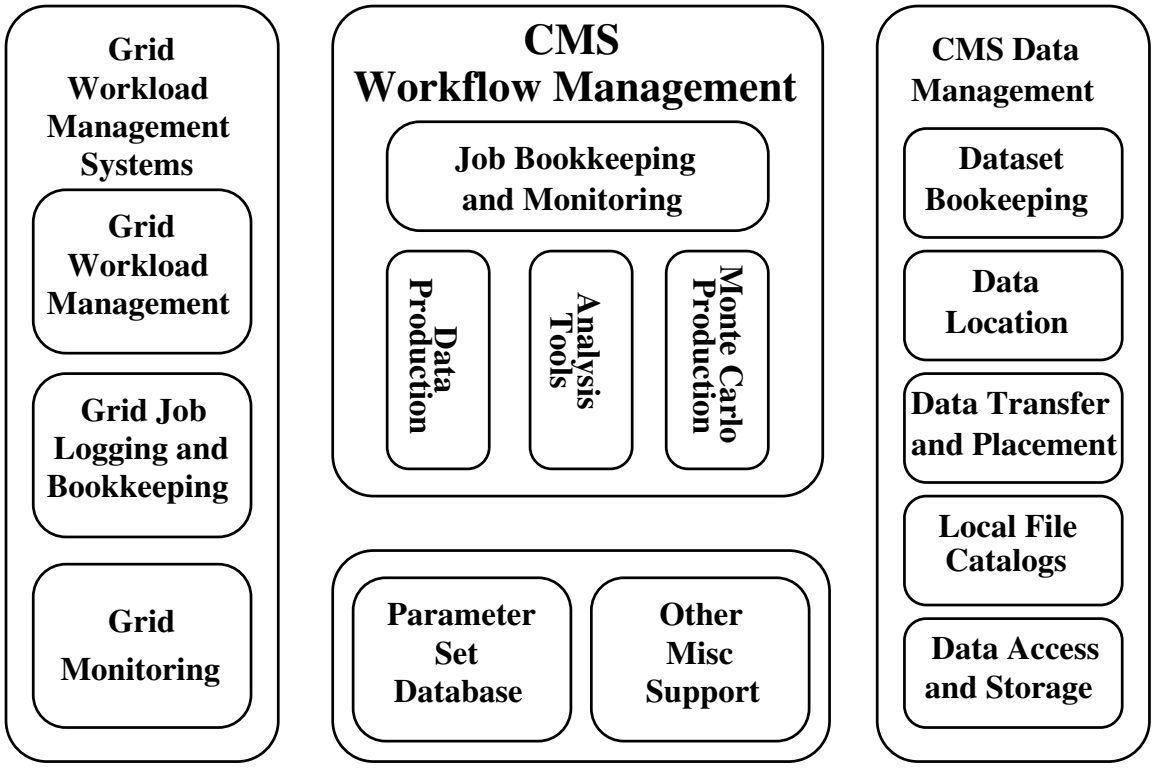


Figure 4.1: Overview of systems and services supporting the CMS workflow management system.

implemented on top of or interfacing to standard WLCG-wide services and interfaces.

Fig. 4.1 shows the overall architecture of the CMS computing system along with the most important systems and services:

- Grid Workload Management System - the core Grid systems and services allowing CMS to make use of and access distributed computing resources
- Data Management - the core services which provide the tools necessary to manage the large amounts of data produced, processed and analysed by the CMS computing system
- Other CMS-specific services - other services supporting specific needs of CMS applications and software

Tying all of these pieces together into a coherent system supporting CMS physics is the CMS Workflow Management system. This system will support all necessary workflows: data (re-)reconstruction, calibration activities, Monte Carlo production, AOD production, skimming and general user analysis. It will also shield the users and operators of these systems from the full complexity of the underlying systems and services.

In subsequent sections of this chapter we will examine each of the component systems and services one-by-one. We will then return to the CMS Workflow Management and work through several important example workflows to demonstrate how these components will work together and how these will be used on the distributed computing system.

4.4 Data Management System

4.4.1 Data Organisation

The computing system needs to support both physicist abstractions, such as "dataset" and "event collection", as well as physical "packaging" concepts native to the underlying computing and Grid systems, such as files.

We define an "event collection" as the smallest unit that a user is able to select through the dataset bookkeeping system described below, i.e. without using an analysis application which reads individual events. An event collection may correspond to the event data from a particular trigger selection from one given "run", for instance. We don't impose here any restriction that these event collections be produced with the standard Event Data Model (EDM). These could very easily be any type of user-defined "ntuple".

We generically define a dataset as any set of "event collections" that would naturally be grouped and analysed together as determined by physics attributes, like their trigger path

or Monte Carlo physics generator, or by the fact that they represent a particular object model representation of those events (such as the RAW, FEVT, RECO and AOD data formats).

In addition to datasets defined by the production systems we expect that users will want to define their own subsets of these datasets by selecting on criteria such as run ranges, data quality flags, etc. and that these user-defined datasets will need to be tracked.

An important concept to note is that in this model, event collections are the principal job configuration concept used at application run-time, while dataset as a concept is only used by the physicist prior to job submission.

Behind the physicist view of datasets and event collections, the event data will be organised into files that can be handled by the storage and transport systems. We expect that the event collections will in general map to one or more files and that there be some easy means for the framework application to know which files to open given the name of an event collection.

The packaging of events into files will be done in such a way that the average file size is kept reasonably large (e.g. at least 1GB) in order to avoid a large number of practical scaling issues that arise with storage systems, catalogues, etc. when very small files are created and need to be tracked. This doesn't imply that small files cannot be created and handled in a transient way, for example as the output of individual jobs, but that the data production systems will be expected to include "merge" steps in their workflows in order that the files tracked by the DM system in the long term are of adequate size.

In addition to "files" as a unit of packaging, we introduce an additional concept: the "file block". This is just a set of files which are likely to be accessed together. As will be described in subsequent sections in this document, we expect that it will be convenient to group data in "blocks" of 1-10TB for bulk data management reasons. Any given file will be assigned immutably to a single unique file block and global replication within the system will likely be done by file block rather than the single file.

Non-event conditions and calibrations data will be accumulated in addition to the event data and stored separately. In the following, we will call "Conditions data" all non-event data required for subsequent data processing. It encompasses:

1. Detector control system data (DCS) - slow controls logging
2. Data quality/monitoring information - summary diagnostics and histograms
3. Detector and DAQ configuration information used for setting up and controlling runs, but also needed offline
4. Traditional calibration and alignment information

Calibration procedures determine (4) and some of (3), others have different sources. For non-event data which needs to be accessed by applications reading event data, some time-

based key will be used (either the event time itself or perhaps a run number) and that this key will be stored with the event data itself.

4.4.2 Data Management Overview

The CMS Data Management (DM) system is intended to provide the basic infrastructure and tools allowing CMS physicists to discover, access and transfer various forms of event data in a distributed computing environment. The range of use cases includes large-scale data and Monte Carlo production, analysis group productions and individual user analysis and thus the system will provide functionality covering the full spectrum from that needed by large Tier-1 sites to that needed by a single user on a desktop.

As part of its computing model, CMS has chosen a baseline in which the bulk experiment-wide data is pre-located at sites (by policy and explicit decisions taken by CMS personnel to manage the data) and thus the Workload Management (WM) system need only steer jobs to the correct location. Initially the DM system will focus on bringing up this basic functionality, however hooks will be provided for more dynamic movement of data in the future (e.g. by the WM systems in response to the needs of the ensemble of submitted jobs).

The DM architecture is based on a set of loosely coupled components which, taken together, provide the necessary core functionality in a coherent manner. By keeping the components loosely coupled we insure that the future evolution of each part can remain relatively independent, including the possibility that the implementation of any of the components may be replaced entirely if necessary or desirable. These components will be described in subsequent sections of this document.

4.4.3 Data Management Architecture

The basic data management architecture consists of the following components:

- **Dataset Bookkeeping System** - Answers the question "Which data exist?"
- **Data Location Service** - Answers the question "Where is the data located?"
- **Data Placement and Transfer System**
- **Local File Catalogue(s)** - Site local information about how to access any given Logical File Name (LFN)
- **Data Access and Storage System(s)** - Provides access to files, including (perhaps) internal management of replication in a disk cache and/or tape systems

Taken together these provide a coherent set of DM services.

Data management happens at many levels, from the global bookkeeping of the experiments datasets to the moving files to a local environment like a desktop. We define generically two "scopes" for data management: *global* and *local*.

By the *global* scope we mean the system used by CMS physicists to find data which is available to all collaborators. This data will in general be located at Tier-1 and/or Tier-2 sites and will usually be archived in such a way that its long time permanence is guaranteed. Standard contact points for discovering data managed in the global scope will be known and available across the collaboration.

By a *local* scope we mean data created, stored and/or used in some context for which it is not generally available to the collaboration at large. This will likely include both "private" data (which may never be made available to the wider collaboration) and also production data which has yet been published into the global scope. For data managed in a local scope it is not expected that the access points be known to the collaboration. There will likely be a number of instances of local scope data management systems.

Data managed in the global scope will require the full set of system components listed above. For data managed in a local scope, it may be possible that a reduced set of system components is needed. In general it will be possible to "publish" data from a local scope into the global one and tools will be provided to facilitate this. To facilitate the maintenance of the provenance information in a local scope, it will also be possible (for example) to import limited slices of information from the global scope into the local one. The global/local division will for example allow analysis users to create their own user data in a local context, validate it, and then publish it to the global scope for use by their collaborators.

4.4.4 Dataset Bookkeeping System

The Dataset Bookkeeping System (DBS) will provide a standardised and queryable means of describing the event data. For the physicist it will be the primary means for "data discovery" and must answer the basic question "Which data exist?". In particular it must express the relationships between "datasets" and "event collections" as well as their mapping to the packaging units of "file blocks" and "files". The information available from queries to the DBS will be site-independent.

More specifically it will be possible for the analysis or production user to perform queries against the DBS in order to:

- Select existing named datasets
- Create new datasets by selecting subsets of event collections from an existing named dataset

The criteria by which datasets and/or event collections are selected may include run ranges, available "data tier", software release used for a processing, data quality flags, etc.

Once selected, it must be possible to resolve those datasets into lists of event collections for use in configuring analysis or production jobs. Similarly the DBS will also provide the mapping from event collections to the "packaging" units like files and file blocks.

A variety of other information will also be tracked in order to provide the complete description of the event data. In particular this must include data provenance. We define the provenance tracked by the DBS as the following:

- Which event collections are derived from which other event collections
- Which application, software release and parameter set were used to do the transformation (and possibly also a "tag" indicating the conditions/calibrations versions used for the processing)
- (Optionally) which datasets are derived from which other datasets

Certain types of more detailed provenance information, such as which parameter sets were used to produce particular data items *within* an event (e.g. when partial reprocessing is being done, with some things recalculated and others just copied from input to output) will probably not be available explicitly from the DBS. To access this level of detail it will be necessary to open the event collection(s) with some application (e.g. framework-based or ROOT).

We expect that other types of basic information will be valuable in the context of the DBS, either as "summary" information or as part of data discovery and selection:

- Rough estimates of integrated luminosity - The most exact estimate of luminosity may require in addition the use of information stored outside the DBS (in the conditions database, for example).
- (Possibly) information on luminosity "run segments" - These finer grained subdivisions of a run, each with a known integrated luminosity, were found to be a useful concept at CDF/D0. This information may be needed for job configuration in order to insure (for example) that individual jobs run on integral numbers of run segments.
- Information on "runs"
- Data quality flags associated to either "runs", "event collections" or perhaps even whole datasets

It should be noted that there are things which the DBS intentionally will not track, such as detailed job bookkeeping, Monte Carlo requests and their characteristics, etc. These

things belong to the internal bookkeeping of the WM/Production systems. This clean decoupling between the two types of information has a number of advantages including:

- It allows multiple types of data production systems (Monte Carlo simulated data, real data, analysis production, etc.) to evolve independently with minimal coupling to the subsequent consumption of the data that they produce. The coupling is effectively only through the data description required by the DBS.
- It allows many details which are irrelevant for subsequent consumption of the data to be dropped. This could include:
  - Details of how data was split for processing in cases where it was merged by the production system before publishing.
  - Details about data produced by failed jobs which was never made available to the user
  - Details about the internals of any multi-step process being managed by the WM system when intermediate data results are thrown away

It is expected that the DBS will be usable in multiple "scopes". The most global scope will be CMS-wide data. This may be implemented (for example) by a centralised database at CERN with read-only mirrored copies in other locations to optimise access. In addition, we expect that the DBS tools will support the creation of DBS instances with a more "local" scope such as a group working at a university centre or an individual user. This "local" DBS scope will allow for the production of private data, not available to the general CMS user, but the tools will support the eventual publishing of such data to the "global" DBS scope (presumably accompanied by movement of the data to some location where the general CMS user can access it, such as a Tier-1 or Tier-2 centre).

The DBS is VO specific in that it is keeping a description of CMS specific data structures using an application specific schema, where the primitives are not the raw Grid files, but rather the more complex CMS datasets and data blocks. Hence it is an application service that will be constructed by CMS on top of Grid services and database components.

4.4.5 Data Location Service

The DBS described above simply allows a CMS user to determine which data exists without regard to where replicas of that data may be located in the distributed computing system.

To find data replicas in the global scope a separate Data Location System (DLS) is used. This system maps *file blocks* to storage elements where they are located.

It is likely that the DLS will be implemented as a 2-tier system. A local DLS instance present at each site publishes the data-blocks available locally while a global index aggregates information from the local DLS instances. Queries placed to the global index result in a list of sites that have a given data-block. Information is entered in the local DLS index, for sites where the data is located, either by the production system at the end of the production or by the data transfer agents (see below) at the end of transfer. In both cases a data-block is only publicised at a site when complete. Site manager operations may also result in modifications to the local index (e.g. in case of deletion or loss of a data-block). Access to local data (i.e. within the same site, in a non-Grid fashion) never implies the need to access to the global DLS: if data are found to be present locally (e.g. on a personal computer), they are directly accessible.

The DLS will also provide some means for expressing the concept that certain replicas of data are considered *custodial*, i.e. that the experiment considers that copy of the data at that site to be permanent. Sites take custodial responsibility (through Service Level Agreements) for copies of the data, that cannot be removed without insuring with the experiment that either the data is no longer needed or that some other site takes on the custodial responsibilities for the data. (It will be Tier-1 sites that take custodial responsibilities for data.)

In addition, in case the underlying SE system provides data access cost estimation (e.g. whether the data-block is normally on disk or on tape), this information may be exposed to the outside through the local DLS. This is non-baseline and is not intended to expose the real time state of the disk cache, but rather the general policies for blocks of data.

The DLS may be provided by suitable modification or evolution of existing general Grid components.

4.4.6 Local file catalogues

The DLS described above does not provide physical location of constituent files at the sites, or the file composition of data-blocks. They only provide names of storage elements at sites hosting the data. The actual location of files is only known within the site itself through a Local File Catalogue. This file catalogue will present a POOL interface which returns the physical location of a logical file (known either through its logical filename which is defined by CMS or through a Global Unique IDentifier, GUID). CMS applications only know about logical files and rely on this local service to have access to the physical files. Information is entered in the local file catalogue in a way similar to the local index of the DLS, i.e. by the production system, by the data transfer agent or by the local site manager.

Note that if the local SE itself has an internal catalogue and the CMS Logical File Name (LFN) namespace is overlayed on the SE namespace, the functionality of the local file catalogue may be implemented by a simple algorithm that (typically) prefixes the logical

file name as known by the CMS application with a site-dependent access prefix that is provided by the local configuration. In this case the local file catalogue is effectively integrated into the SE itself and no extra information needs to be entered in a separate file catalogue when files are added or removed from the SE. This is the case for instance when data are copied to a personal computer (e.g. a desktop) for iterative analysis.

It is expected that these local file catalogues will be able to provide attributes and information about the files (e.g. checksum, filesize) but will not contain CMS-specific attributes describing the file content.

The file catalogues are expected to be at least as robust as the data storage itself and to sustain very high-scale performance.

4.4.7 Data Placement and Transfer System

Overview

As described elsewhere in this document, the core infrastructure for data transfers and placement is formed around a relatively stable set of storage services offered to CMS at Tier-0, Tier-1 and Tier-2 sites. A number of Tier-2 and Tier-3 sites form a more dynamic infrastructure around these larger, more stable sites.

The data placement system is used to define, execute and monitor experiment policies on where experiment data is to be located. This layer manages allocation and release of storage resources as well as data transfers at the level of datasets and file blocks.

Replicating and moving individual files is handled by the data transfer system. It handles reliable background transfer of files from multiple sources to multiple destinations at maximum possible throughput. It provides estimates on latency and transfer rate for scheduling purposes. The transfer system is aware of file replicas during transfers, but once data placement and location systems have been notified and transfer details have been archived, drops the knowledge about the details. The transfers operate largely asynchronously and separately from the other data management components; it is not required for files to be known to the other components for transfers.

The data placement and transfer systems are implemented by the PhEDEx project.

Functional requirements

Managed and structured data flow. CMS requires a data transfer system with a global view of a transfer topology. All sites in the system do not connect to all other sites for data transfer. For example, not everybody can connect to the Tier-0 at CERN, yet the connectivity is quite free among and below Tier-2 sites. We require the system to have a global view of transfers that pass through several storage systems, such as tape to disk to disk to tape. We also need to be able to efficiently release files at disk buffers, especially at the Tier-0 but also for Monte Carlo production at Tier-2 sites, which requires knowledge

about whether files have made it to the final safe destination, not just the next transfer outward – for instance raw data to the Tier-1 tape storage.

Multiple transfer modes. Data transfers are initiated for a number of reasons. The detector facility continually produces data and must *push* it to a number of destinations. If a destination is unable to accept data at the rate the buffer is filled up, it must be possible to automatically retarget the data to another destination – and still transfer it to the original destination after recovery. Simulated data must be delivered in similar fashion to a Tier-1 centre for custodial copy. Both of these are examples of *stream* mode push transfers, although a push can also be carried out for a specific set of data in one-off operation. In addition, it must be possible for sites to *pull* data they are interested in. In general it is to be assumed that once such a data *subscription* is known, especially for "infinite" primary datasets, transfers will take place autonomously without continuous operator attention.[1]

Multiple priorities and scopes. The system must be able to address possibly conflicting or competing priorities for the collaboration as a whole, individual physics groups and ad-hoc groups and individuals. It must be able to merge fairly global and site-local priorities in line with experiment policies. It must be able to scale for collaboration-wide transfers as well as those made by an individual user to a personal computer.

Other requirements. The data transfer system is to be a transfer system, not a replica catalogue, nor is it expected to take care of the overall "workflow." As such, it is expected to scale by number of files in transfer, not the total number of files. It should be able to transfer any file from anywhere in CMS to anywhere else.

Specific sites will need their own policies to manage the efficient movement of data into and out of the site, and to serve it for analysis. This will include being able to delete files to reclaim space, either automatically when the space is needed for other purposes or on demand when a site is being reconfigured in some way. The data placement service will be able to know if a given dataset is needed at a site where it still exists, or if it is already safe elsewhere. It will be able to trigger automatic garbage collectors once a dataset or file block is known to have been completely moved between locations, or will be able to produce a list of files which can be purged at any time the "site-admin" wishes, within the constraints of the CMS policy.

System requirements

The data placement and transfer system must be robust: the malfunction of any one part should only affect its immediate neighbours, and no one part should be able to bring the whole system down. The transfer system itself should consume limited amounts of resources and should be easily able to saturate any network or storage infrastructure given to it. The transfer system should operate autonomously from other data- and workflow management systems, production operations, worker nodes, and so on.

[1] In this context "push," "pull" or "stream" describe the operation logically, not how the transfer is technically to be executed.

We expect that CMS storage systems will present SRM [17] interfaces, and SRM is the strategic choice for storage management interfaces in the WLCG. The transfer system in addition supports commonly used protocols, in particular GridFTP, as we expect to interface to non-SRM interfaces at some smaller sites. At the extreme, transferring files to or from a local computer may access local files directly.

Data placement system

The data placement system keeps track of files, generally in form of blocks, and data placement requests: open-ended subscriptions and one-off transfer requests. A placement request specifies the data to be transferred, priority, and whether the resulting copy is a custodial one. This is in effect the execution of the experiment data placement policy, including the option to divert files to a fallback destination if they do not reach the primary destination quickly enough.

When new files are created, the data placement system is informed about the file and where it is available. It then manages the creation of actual file transfer requests, monitors transfer progress, and provides means to notify sites and other systems about the progress made, for instance to publish completely transferred file blocks in the data location system.

The data placement system provides a means to identify replicas to be "released" when outstanding transfer requests have been fulfilled. There is no means to move files specifically, as it is not easy to define a meaning of a move for a file with several replicas, any of which could be selected for transfer.

Data transfer system

The data transfer system operates at file level and autonomously of other components. Any file can be transferred, not just event data; files registered for transfer need not be known to other components prior to transfer. The transfer system does however assume the files are immutable and that within experiment policies it has the freedom to select the best replica when making transfers. The transfers are made asynchronously in the background, the baseline system does not make transfers in response to file access from jobs on worker nodes.

The transfer system typically receives transfer assignments from the data placement system, including a file, destination and priority plus where replicas are known to exist for the file.

Transfer assignments can also be created directly in the transfer system, for instance by the WM output harvesting when output file destination is defined in the job description. Such requests would typically be executed at low priority and the files might not be made known to the data placement or location systems.

The data transfer system uses the concept of a *storage overlay network* in which nodes are disk and tape storage systems and edges are possible transfer links. The system takes care of reliable end-to-end transfers in this overlay network, factoring in experiment transfer policy and priorities.

Related components

Data transfer and placement systems interact heavily with the data storage and site storage resource management systems, for the latter typically a SRM. However the actual transfers are almost always executed as third-party transfers, so the transfer system itself does not require significant amounts of network or processing capacity. Grid file transfer tools are used for the actual transfers.

The data placement or transfer systems do not interact directly with the data bookkeeping system. In general the workflow or production bookkeeping systems inform data bookkeeping system separately about new files without reference to where those files might be available. Bridging agents are used to cross such workflow steps.

Neither the data placement nor the data transfer system form a replica management system as introduced to date by Grid projects. The data placement system manages replication at the level of the larger blocks, not for individual files. The data transfer system only knows about replicas while transfers take place, but does not for instance know file paths at different sites. File paths are only accessible within each site, not outside them. There is no transfer of files on-demand by worker-node jobs. There is no central catalogue that knows about every file at every location with a replica.

Beyond baseline

Beyond the baseline we are considering a number of options. Priority and policy management is likely to require improvements over time, in particular to provide hard and soft deadlines for transfers and better management of bandwidth usage, plus moving towards more distributed policy management. The file routing will most likely have to adopt techniques from more advanced IP routing algorithms, in particular to distinguish and properly handle common congestion and error situations. Transfers probably should be able to tune automatically at small scale local low-level parameters such as degree of parallelism.

Finally, we are researching making the transfer network more dynamically distributed to facilitate the participation of transient and small nodes, such as personal computers. A related change is to allow agent configurations to be changed remotely and dynamically.

4.4.8 Data Access and Storage Systems

We describe the baseline storage systems that sites will have. CMS systems interface to site storage from the Grid side through the Data Placement and Transfer System, possibly through a layer of file transfer services or directly through the SRM [17] storage management interface. CMS applications running in jobs will interface to storage through a POSIX-like interface, where file-open commands may require the specific syntax of Storage URL's (SURLs).

Storage systems have an internal catalogue (or even just a file system) that implements a

local namespace. The use of SURL addressing allows an abstraction of physical storage (like actual disk mount points etc). A site, in particular a very small site (e.g. desktop), may expose physical layout, but takes on full management responsibility then.

The site storage system interface from the Grid side for all of the larger centres will be SRM. A detailed list of SRM functionality is being worked out, but will likely include transfers, space management, advisory delete, etc.

Sites, and in particular Tier-1 sites, will provide storage systems technically capable of providing long-term, custodial storage of CMS data. For this responsibility we expect that sites will make Service Level Agreements specifying the availability, throughput, error rate etc.

Given current technology predictions we expect that these sites will require tape libraries as secondary storage, with a large disk-array as a front-end to allow applications to access data in direct-access fashion. We will outline the required storage and data access primitives and thus will allow storage system providers to optimise their file management (e.g. large classes of event data files will never be modified or have data appended to them once they have been written the first time).

We foresee that CMS will provide and require a site to track file attributes like check-sums, data access, etc. We also envision to provide information about allowed access patterns to specify which files are read-only.

For Tier-2 sites CMS will allow for more lightweight storage systems. These will be used in particular for placement of datasets used for analysis that can relatively easily be replaced through re-generation or reimport from storage systems at Tier-1 sites.

In addition, we will require additional storage for physics analysis use at sites. This storage space will need to implement the full spectrum of POSIX IO, in particular re-writing records etc, for user storage of ROOT trees, for user code and libraries etc.

4.4.9 Conditions data

All Conditions data required by the online HLT event filter and all offline event processing applications are stored in one conditions database (condDB), at least at conceptual level.

The CMS current baseline is to host the Master condDB as part of the Online system (ORCON). ORACLE is the candidate technology of choice. It will be accessed directly only by online applications such as high level trigger and data quality monitoring tasks. Data required for offline applications (simulation, calibrations, reconstruction and analysis) will be replicated in an offline condDB (ORCOFF) whose master copy will still be hosted by the Tier-0 centre. Further data distribution to T1s and eventually to T2s may use ORACLE replication technology or custom solution. The final data access from client applications will make use of a system of distributed caches along the line implemented

by commercial systems such as Oracle (Data-WareHouse and Web-caches), open-source software such as Hibernate or custom solutions such as FronTier. [18].

4.5 Application and Job System Services

4.5.1 Parameter Set Management System

The CMS application framework can be configured and can describe its configuration using a *Parameter Set* system. The Parameter Set system is responsible for defining in an unambiguous way all of the input parameters for the individual algorithms included in the applications. Together with a tagged version of the software algorithms and perhaps a conditions/calibrations tag, the Parameter Set ensures a well-defined provenance for the output produced by the application as described in section 4.4.4.

The Parameter Set system will be the only way to define application parameters, no other mechanism will be allowed. As a prerequisite to run an application, all the Parameter Set which will be used by the application itself must be defined and registered with a mechanism provided by application framework. The baseline model for the use of these Parameter Sets is that the set of all Parameter Sets used to produce data will be tracked in a database, the Parameter Set Database. The application framework itself will be capable of writing the ensemble of individual parameters into this database, providing a globally unique identifier for each Parameter Set.

The details of how these Parameter Sets are used internally by the application framework itself is outside the scope of this document, however the management and bookkeeping of the Parameter Sets is part of the computing system.

The parameters inside the Parameter Set are divided into two categories: untracked and tracked. The former subset comprises all those parameters which change the run time behaviour of the application, but not the results: *e.g.* verbosity level, debug option, etc ... and the tracked parameter, those which really matter for the provenance, and will be defined before the job execution by the user. The unique identifier will insure that two jobs configured with identical sets of tracked parameters will be seen as having the same provenance.

This parameter set database will be used by the Workflow Management system described below to configure jobs. The parameter Set needed by the application will be extracted from the database and put into file, which will then be shipped together with the job: this file will be part of the job configuration. No access to the Parameter Set database will thus be necessary from individual worker nodes: all access to the Parameter Set database will happen from the User Interface prior to job submission.

As all data described by the DBS should have a well defined provenance (see sec. 4.4.4,

every Parameter Set ID referenced by the DBS must correspond to an entry in the Parameter Set Database. Note that this also implies that the same global/local "scope" structure described above for the DBS will also exist for the Parameter Set Database. There will likely be different back-ends for the different scopes: for a generic user, file based (*e.g.* SQLlite), whereas for the global CMS scope of official data production, a database management system will be used.

4.5.2 Job Bookkeeping and Monitoring System

The CMS WM tools described later in this chapter rely on a job bookkeeping and monitoring service. This service provides the basic infrastructure for tracking the overall status of individual jobs as well as real-time monitoring (when possible) of the progress and resource utilisation of the jobs. CMS expects to use a VO-specific job bookkeeping and monitoring system (BOSS [19, 20]), that knows about the CMS job structures and configurations and is tailored for these needs.

The bookkeeping aspects of BOSS provide the means to log and track all user jobs, to keep control about what is happening and about what happened to ensure provenance and reproducibility. It logs all information, either related to running conditions or specific to the tasks they performed, in a database accessible from the UI from which the job submission takes place.

The logging will include information which is relevant regarding the job configuration, interaction with WMS, history of execution etc. It thus contains much more information than what is needed to eventually populate a local or global DBS for future data discovery. A suitable tool will be provided to extract and summarise all logging information.

The system is capable of referencing the grid logging and bookkeeping system so that the user will be able to access the grid system transparently. Information such as the running status of the job and eventually its exit code must be available through this channel.

Application specific information is obtained either by parsing the job output (a non-invasive mechanism), or via direct communication from an instrumented user application. Recently *log4cplus* [21] was chosen by the CMS software group as the only logging mechanism for applications being built on the new Event Data Model. The system will be able to treat the information produced through this channel.

The job bookkeeping database backend may vary depending on the environment. For analysis users it will in general be lightweight and file based (*e.g.*SQLlite), with no need of a dedicated server. For larger scale "production" users a more robust and scalable database management system will be used.

In general information is available in the database and to the user only at the end of job execution, i.e. when the job output is made available to the submitter. If the job output is not made available to the submitter the job is considered to be failed. Thus,

by definition, retrieving the logging information together with the job output is reliable. This is considered as the baseline solution even though other more scalable mechanisms will be possible. The client program will be able to transfer the logging information to the database transparently to the user.

An additional service provided by the job bookkeeping and monitoring service is real-time monitoring. This optional service would make the job information available in the afore-mentioned database while the job is still running. The information is collected on the WN as it is produced by the job and sent to a dedicated service (Real-time monitoring database) by a suitable monitoring plug-in that is executed in parallel with the user program. Eventually the information stored in the real-time monitoring database is accessed (with a POP-like mechanism) by the same client running on the UI that is used to access the information stored in the bookkeeping database.

Failures in the functioning of the monitoring plug-in will not affect the normal running of the user program nor the availability of the logging information in the bookkeeping database at the end of the job.

Options for the implementation of the real-time monitoring service depend on the assumptions on the availability of outbound connectivity from the WN's or of suitable tunnelling mechanisms (e.g. HTTP proxy, R-GMA servlets, etc.) on the CE's. In general it is expected that the same real-time monitoring database will serve many submitters and is provided on some highly available server at a Tier-1 or Tier-2 centre.

4.6 Software Packaging and Distribution, Configuration Management

CMS software and externals used by it are distributed in the form of RPM packages [22,23] such that it is possible to install both a full software development environment and only the parts required at run-time [2]. CMS requires to be able to directly verify that the advertised software is installed correctly, by checking for end libraries and programs mentioned in a package manifest and by verifying file checksums. CMS also requires to check the system software configuration in a similar fashion.

The software is installed at each site under a single root location in a hierarchy defined by CMS. The location must be accessible from all worker nodes on the site and easily discovered, typically via single environment variable. The location must be writable by the CMS software installation managers. The area must be reserved for CMS software and should not include software from the underlying system nor other experiments or projects.

[2] This is so that it is not necessary to install full development environment and documentation on systems which will only need to run the software.

In general, the same software is assumed to be accessible to the developers and usable as SCRAM base project areas. In the baseline system CMS does not expect to be able to compile software on the fly on the worker nodes – the code will come prebuilt. User code is supplied as a prebuilt custom code based on preinstalled public CMS software, and is delivered directly in the job sandbox.

The RPMs are provided through a single central authoritative software repository. Packages from this repository may be replicated to other repositories either using the data transfer system or by other means. A site must provide for automatic installation of software by some combination of submitting Grid jobs and deploying an automatic software installation service appropriate to the Grid involved.

Information about installed software should be advertised in the Grid information system, for use by the workload management for job matching. A site can remove software, provided it also removes the corresponding entries from the information system. The CMS software managers have the ability to remove software by submitting a Grid job to the site.

4.7 Grid Workload Management Systems

CMS expects the WLCG and sites to provide a Grid Workload Management Systems (WMS) that has certain characteristics, as described in this section. We anticipate that different implementations of Grid WMS will exist and be used by CMS in the different Grid worlds (EGEE, OSG, NorduGrid, etc.).

This section describes the CMS expectations on and requirements for minimal functionalities of these systems by specifying a basic "reference" architecture. We also provide some performance metrics and briefly describe the status of emerging systems within EGEE and OSG. We conclude this section with our understanding and desires for interoperability between Grid middleware deployed by the EGEE project and OSG in terms of a set of baseline services of the WLCG.

4.7.1 Basic Architecture

The basic functionality of a Grid WMS is to schedule jobs onto resources according to the VO's policy and priorities, to assist in monitoring the status of those jobs, and to guarantee that site-local services can be accurately discovered by the application once it starts executing in a batch slot at the site. In the present section we define our expectations and describe a basic architecture that we expect from a Grid WMS.

4.7.1.1 Job Prioritisation

CMS requires the WMS to provide functionality to resource providers and to the CMS VO, to adjust relative job priorities for implementing both VO and site policies. CMS needs the ability to set policy and priorities concerning the usage of resources pledged to the experiment at a given site, and the WMS should provide such functionality.

CMS requires the ability to define which users or group of users have precedence over others, with as much flexibility as possible. CMS needs to be able to define and tune the priority ranking among $\mathcal{O}(10)$ analysis and production groups with a latency for putting those policy changes into practice of not more than a day.

At the CE level, the owner of the resources is allowed to define the load balancing between different user organisations, e.g. CMS, ATLAS, LHCb, ALICE, etc., as well as their local users. Beside that, CMS administrators must be able to decide easily how the resources available to CMS should be shared among the various groups, sub-groups, and users within CMS. A first prioritisation balancing will be done typically between official production (both data and MC) and analysis. Furthermore, balancing among different analysis groups must be supported, according to experiments priorities and milestones. Within a given analysis group it must be possible to guarantee fair share access for the CMS users within that group. It is highly desirable to be able to delegate control over a group's access to resources to personnel in that group. The groups themselves would thus decide and specify allocation of resources to their sub-groups or even individual users. However, this will not be in the baseline.

There are several ways to provide such functionality: in any case the ability to define on a Grid environment groups of users and roles for user is a key element. An ideal solution would then allow modification of the configuration of the Grid scheduling mechanism at different granularity depending on the roles attached to the user's proxy. The lowest level of privilege would then allow only the reprioritisation of ones own tasks previously submitted to the WMS. This would avoid a user getting stuck with a small high priority task behind their own previously submitted large low priority task. As user priorities often change on a day-by-day basis, this functionality is essential to guarantee high level of user productivity.

4.7.1.2 Baseline workflow

The baseline solution for a Grid WMS is shown schematically in fig. 4.2. The WMS acts as an distributed scheduler with matchmaking capabilities, not different from any advanced local batch scheduler.

The Computing Elements (CE) declare themselves available and publish information about their status and resources, which can change dynamically. The WMS is the connection between the user, which works on the User Interface (UI) and the CE. The UI is able,

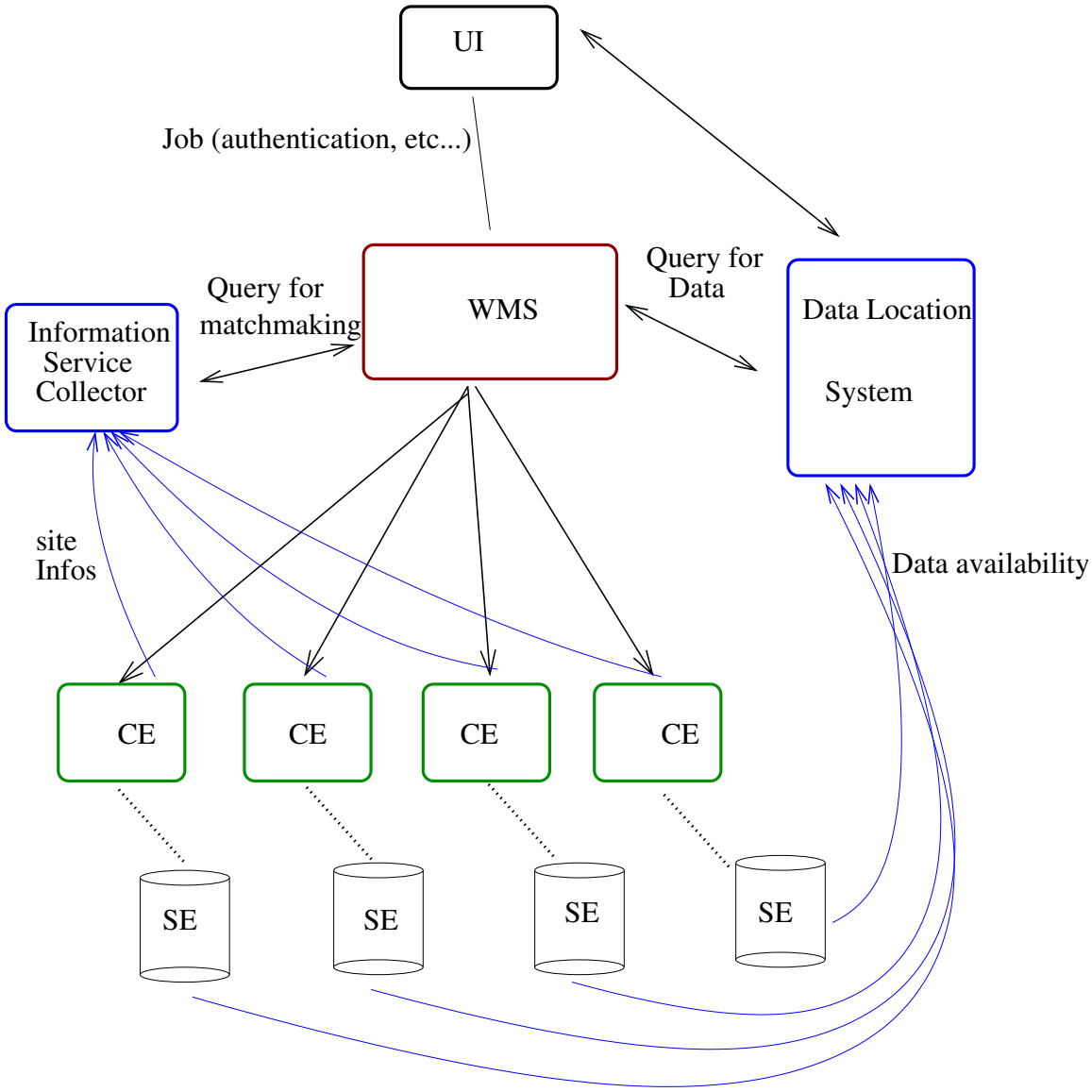


Figure 4.2: Schematics showing the baseline WMS architecture.

via proper middleware and authentication, to submit jobs or task to the WMS. The user can specify a set or requirements which the CE must fulfill in order to be eligible to run the job. As part of the requirement, the data location is one of the most important. This can be accomplished in two ways. The data location can be done at UI level and then the job is sent to WMS with the requirements that the CE has access to the Storage Element (SE) where the data is located. Alternatively the data location is done directly by the Grid WMS, so the requirement is simply to route the job where the data is available. The requirements are matched against CE published resources by the WMS, which then choose, using a ranking algorithm, the best suitable CE where the job will run, and then, the WMS route the job to the CE. The local batch scheduler on the CE will, in turn, take care of the job and execute it on a Worker Node (WN).

The major advantage of this approach is the simplicity, together with the capabilities of more advanced use thanks to the match making functionalities of the WMS.

4.7.1.3 Beyond the baseline: Hierarchical Task Queues

While the baseline system described above should be sufficient for most purposes, it has some disadvantages. An automatic mechanism does not exist to check whether a CE (or even a WN) has a problem that would prevent it from successfully executing a CMS application. Examples are problems with the software configuration required for the application or problems with data access. Currently the only way to spot these or other problems is to try to run an application and fail. Thus, only a efficient monitoring mechanism and responsive action by administrators can provide fast solution. Also, more advanced functionalities can be complex to implement. For instance, data placement or movement on–demand, which can be the basis of efficient use of opportunistic resources, or jobs submission automatically triggered by appearing of data somewhere (the so–called real–time analysis), etc ... We thus foresee a possible beyond-the-baseline solution.

Figure 4.3 depicts a schematic of the hierarchical task queue architecture that we envision to be implemented by the Grid WMS. The distinguishing features are a couple of task queues that are controlled by CMS policy, as well as an agent that harvests batch slots (HBS) from the site batch system for use by CMS. The "CMS Batch" agent that is submitted by HBS calls the local task queue when it is first launched by the site batch system, as well as anytime it finishes an assigned job. The local task queue assigns jobs according to its local policy. As the fill state of the local task queue goes below some watermark, it requests more jobs from the global task queue. The global task queue assigns jobs to the local task queue according to its policy. The global policy implements site selection based on DLS information or JDL prepared by the CMS WM. The local task queue may implement CMS policy for the site, and communicate it as part of the job request to the global task queue. Jobs are thus "pulled" from the compute node via the local task queue from the global task queue.

The HBS would submit CMS Batch based on the pending status of the single queue for

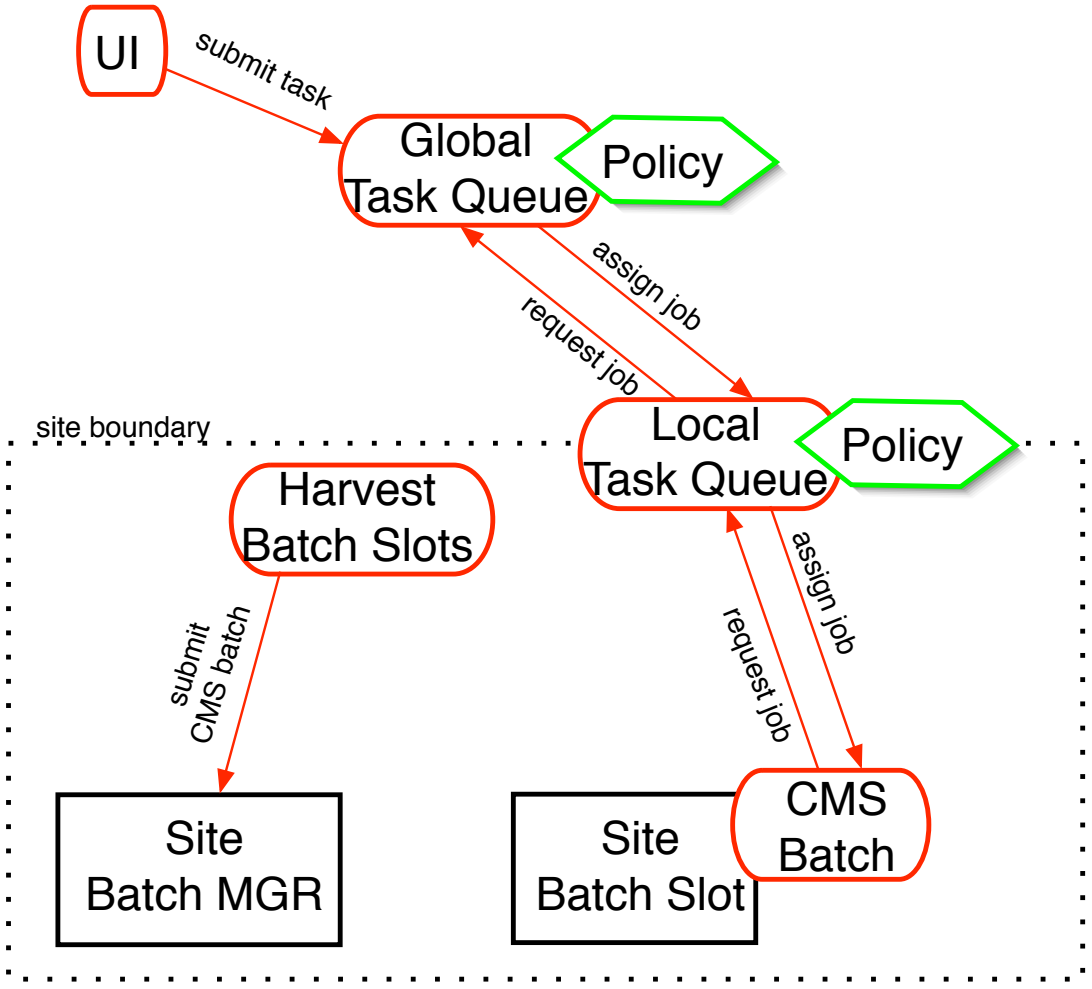


Figure 4.3: Schematics showing the hierarchical task queue architecture.

CMS in the site batch system, as well as a low watermark of the local task queue.

By exposing the local and global CMS policy configuration via a web service that uses VOMS roles for authorisation, it is in principle possible to dynamically change the CMS policies at any site. This would clearly satisfy the requirements for flexibility in job prioritisation expressed in Section 4.7.1.1.

This architecture addresses important requirements, beyond the CMS baseline:

- The compute resource usage policy is fully under CMS control, within the CMS-allocated resources at a given site, and can be adjusted dynamically to reflect the day-by-day priorities of the CMS collaboration.
- It provides a global perspective of the load per data block, and thus allows for simple algorithms to implement block of file replication if the sites that host that block of files are oversubscribed while other sites have spare CPU resources. Such replication is a "beyond the baseline" functionality of the CMS computing system.

- It provides at least short term predictability of the computing resources available at a site, and increases CMS control over execution latency of its workloads. This is especially important in resource poor environments where resources are shared without fixed resource quotas.

- It allows CMS to better use the CPU and wall clock times allotted by the site batch system for each batch slot lease because multiple jobs can be run consecutively within the same batch slot lease. This decreases CMS's dependency on the performance characteristics of the site batch scheduler, and may provide improved operational stability, especially in situations where the site batch scheduler experiences excessive load from local user submissions, for example.

- It allows for a variety of future "beyond the baseline" extensions that might use preemption or suspension of long running production processes in favour of short but high priority jobs. An extreme example might be running PROOF slaves parasitically on CMS batch as described in more detail in Section 4.8.7.

4.7.1.4 Grid Monitoring

Monitoring of the computing resources used by CMS will be critical to allow their efficient use by the experiment. For example, information about data access patterns is needed to optimise data placement by replicating popular file blocks or by pinning on disk data which is accessed frequently. Optimisation of data transfer routing with PhEDEx requires precise knowledge of network and storage resources status. We also expect that monitoring will provide the accounting information on resource utilisation by CMS users and groups necessary to enforce experiment policies on resource usage.

Monitoring also plays a critical role in spotting problems with the system. Making such monitoring information easily available to the users can substantially decrease the support load on the operations teams because users learn to answer their own questions, and diagnose their own problems.

Finally, precise computing resource status is needed to enable the Grid WMS to perform match-making between job requests and computing resources.

We expect that many varieties of monitoring will be needed:

- Computing and storage resources: CE, SE, WN

- Grid specific services: EGEE Resource Broker, information services (such as bdII), catalogues, etc...

- CMS specific services: DBS, DLS, PhEDEx services and UI's where they are running, etc...

- Network status.

Examples of such monitoring systems today are GridIce and MonALISA. Both systems have different strengths. We plan to create a unified CMS view and CMS projection of the Grid, using the Grid monitoring systems, and benefiting from the strengths of the available systems.

In addition to Grid monitoring, CMS depends on it's own job bookkeeping and logging system for information that is specific to the CMS WMS. The Grid monitoring system must allow the CMS job bookkeeping and logging system to relate tasks and jobs as monitored by CMS to the corresponding information in the Grid monitoring system.

4.7.1.5 Site-local services discovery

One important principle is that jobs sent to standard sites where CMS-specific services will be running should be configured in a site-independent way. The framework configuration file should contain only "logical" or "site-independent" information such that the jobs can in principle run at any site to which they are steered. This has several advantages:

- Site local information remains local and thus can be changed as needed by the local sysadmins (in any way consistent with currently running jobs continuing to run and new jobs picking up the new configurations when they start on the worker node)
- The workload management tools are simplified in most cases in that they only manage the minimum amount of information necessary.
- The Grid WMS may decide to send jobs to sites not foreseen at job submit time.

Jobs discover the site-local configuration information after they start on a worker node at the site. This site-local configuration will usually be part of the VO-specific site configuration (in the most primitive form, source some environment script), but this may depend on the environment.

The set of local services CMS currently expects that a job should be able to "discover" from a Worker Node includes:

- A "contact" string for the POOL-compliant local file catalogue
- Local location of CMS software (typically on some shared file system).
- Local access means to conditions data
- The local pointer to the job real-time monitoring server, if available and needed

- Definition of "close" SE and/or a local agent of the data transfer system, to allow for asynchronous handling of job output after the job finishes

In all likelihood, the job will need to be able to discover not just location of services but also type of service. E.g., it is possible that different sites implement different SEs for output data handling, and quite likely that even the functionality for monitoring of the running job differs between sites.

4.7.2 Workflow requirements and performance metrics for Grid WMS

This section lays out the most basic required functionality as well as performance requirements by CMS for the Grid WMS infrastructures on EGEE, OSG, NorduGrid etc.

Functional Requirements:

- Description of job dependencies by Directed Acyclic Graph (DAG) should be supported. In particular, the WMS must have the capability to deal with a flat DAG, often referred to as collections of jobs, *job cluster*, or *task*.
- It must be possible to perform bulk operations, such as submission, status query, cancel, etc ... on a complete collection of jobs, or a DAG.
- It must be possible to perform the same kind of operations also on the individual jobs within a job cluster or DAG.
- The Grid WMS should be able to contact directly the DLS, via a proper interface, in order to perform data discovery.
- While in the baseline the function of CMS job splitting is done on the application service level, as a beyond-baseline functionality job splitting might be implemented at the Grid WMS level.
- The WMS should be able to resubmit the same job to the very same resources where it has been submitted before. This is very important in order to debug problems.
- More generally, the WMS should be able to keep memory of the execution of a job, in order to provide information about where, when and how the job has run.

The last two of these requirements could alternatively be satisfied if the Grid WMS interfaced with the CMS job bookkeeping and logging system to log the site a job was submitted to, as long as the Grid WMS furthermore supports steering of jobs to sites.

Performance metrics

- Submission of collection of $\mathcal{O}(1000)$ jobs should happen within few seconds. It is expected that even a typical data analysis task will translate into submission of $\mathcal{O}(1000)$ jobs.

- the Grid WMS must implement fault tolerance and load sharing in order to guarantee stable operations on a 24x7 basis irrespective of number of pending or running jobs or DAGs.

- The Grid WMS needs to be able to schedule jobs at a rate sufficient to keep all batch slots available to CMS busy. We expect this to require continued development effort by the Grid WMS developers in order to keep pace with both the expanding world wide data Grid as well as the expanding CMS data volume.

- The developers of the Grid WMS must strive to provide infrastructure that is reliable enough that even large tasks can reasonably be expected to complete within no more than two retries of failed portions. This implies that the development team strive for a failure rate of less than the third root of the number of jobs in large tasks. We foresee tasks with up to $\mathcal{O}(10000)$ jobs in 2008. This implies a requirement of job failure rates due to Grid WMS errors at less than 5% of the jobs handled by the Grid WMS. The reliability of the Grid infrastructure is essential for the overall success of the CMS distributed computing system.

4.7.3 Grid WMS Implementations

EGEE WMS Implementation

The WMS developed by the EGEE project is referred to as the Resource Broker (RB). From the User Interface (UI) a user submits jobs directly to the RB. In the CMS baseline solution, where the data location will be done at UI level by accessing DLS, the CMS WM submission tool CRAB [24] will take care to define the SE's in user directive for the job. Since data may be located in more that one place, the RB is still expected to match the best available resource among the possible ones. The RB choice will also be driven by other user requirements, such as availability of a specific CMS software version, length of the queue, memory required to run the job and so on. Since it is required that the RB is able to contact directly the DLS via DLI interface [25], the data location can be done also at the RB level. In this case, the user needs only to specify the data blocks he wants to access, and the rest is transparent to the user.

The RB as implemented in gLite today has some of the features described in Section 4.7.1. However, it is missing an interface that would allow dynamic modifications of scheduling policies at either the global or local level. In addition, the local task queue submits jobs directly into the site batch system. To provide some degree of control over execution latency, the local task queue may be configured such that it submits new jobs to the site batch system only when none are already pending. More generally, the actual gLite im-

plementation of WMS lacks the fundamental functionality to define policies and priorities at a VO level.

CMS requires provision of some interface that allows specification of scheduling policy as discussed in Section 4.7.1. The lack of HBS (Batch slot harvester) in gLite is acceptable for the computing baseline system. However, it is desirable to arrive at a roadmap when such functionality would be provided.

OSG WMS implementation

There is a difference between OSG and the infrastructure deployed by the LCG in where the line of responsibility between the VO and the Grid services is drawn. Provision of a WMS in OSG is fundamentally a US-CMS responsibility. What is described here is thus a CMS program designed to make the most of the computing resources available via the Open Science Grid. As we expect the vast majority of the resources on the Open Science Grid to be accessible to CMS only on an opportunistic basis, we are designing a baseline system that makes little to no distinction between owned and harvested resources.

On OSG, services are structured such as to minimise threshold for participation in the "OSG marketplace" for both users and resource providers. This is accomplished by introducing an architecture that follows the "Me - My friends - the Grid" [26] concept. "Me" is a thin layer of user interfaces on the user laptop or desktop. "My friends" is the bulk of services that are implemented in the combination of the CMS CRAB system and the UI infrastructure, and possibly part of the functionality of the EGEE RB. "The Grid" is the site infrastructure, which in OSG is quite similar in functionality to a site deployed by the LCG project.

To be specific, only the packaging aspects of CRAB reside on the user laptop/desktop. Job splitting, parameter set storing, job bookkeeping and monitoring is all implemented by "my friends" at the analysis centre (U.S. CMS Tier-2 centre or LPC User Analysis Facility) that the user is affiliated with. Submissions are thus guaranteed to be fast from the user's perspective, even in case of a relatively slow Grid WMS.

In addition to the submission interface, the user sees two query interfaces. First is the query for job status that is satisfied via the job bookkeeping system. The functionality here is identical to jobs submitted via CRAB to the RB or elsewhere as this level of bookkeeping is CMS specific. Second is a query that provides read-only access to the user sandbox environment. This provides read access to both file and process space of the user for all running jobs. A set of tools is being developed to fulfill this task [27, 28].

For job scheduling, we expect to implement the baseline architecture as described in Section 4.7.1. We expect to base this on the same underlying technologies as used in the gLite implementation of the RB. It is a possibility that U.S. CMS sites within OSG will deploy the EGEE RB, depends on the functionality it will provide.

For authentication and authorisation, the requirements described in Section 4.7.1.1 are implemented by Privilege Project [29] components. In 2005, this includes call-out at CE and SE, as well as sitewide mapping services. All of the site infrastructure is compliant

with proxies generated via VOMS. We thus expect there to be only one VOMS service for both OSG and LCG deployments. A detailed architecture of these components is described elsewhere [30].

Operationally, we expect to build up the Grid WMS on OSG by incrementally deploying production quality systems with expanding functionality but largely consistent user interfaces. OSG is committed to interoperability with LCG at the site level, and it will be possible to use the EGEE RB within OSG. A decision to do so will depend on its functionality and performance characteristics.

4.7.4 Interoperability between different Grid deployments

The fundamental requirement of Grid interoperability is for the CMS user interface for job submission and basic job monitoring to be identical for all Grids used by CMS. This implies that the user sees intelligible messages when they try to use monitoring functionality that is not supported by the site or Grid that their job is running on.

The earliest point of diversification is thus just behind the user interface. At UI the input data will be located, and thus also the underlying Grid. This will allow to prepare the job and job description in a way which is compatible with the Grid and WMS where the jobs will be run. Most of the job preparation will be in common in any case, the only difference, which can be substantial, will be in the communication with the underlying Grid WMS. This solution require a basic brokering to be done at UI level, in case that the input data is available in more than one Grid. This would make optimal utilisation of resources between the different Grids very difficult unless some sophisticated decision making algorithm is applied. As a baseline solution, a rather simple decision can be made, based, for instance, using the UI location, or a user preference.

For a better optimisation of resources usage, the interoperability of the Grids should be possible at the Grid WMS level, that is, either WMS should be able to submit to any CE. This is possible only if the CEs of the different Grids will deploy interfaces that are compliant with the other Grids WMS, that is if the OSG CE will be seen from an EGEE RB and viceversa. This solution will be both transparent to end users and allow the WMS to optimise the CE choice.

The WMS deployed by LCG is based on the RB, thus, to allow interoperability toward OSG, it must be possible for the RB to submit directly to an OSG CE. The OSG CE deployed in 2005 satisfies this criteria by deploying the "Generic Information Provider" version 1.2. We expect both LCG and OSG to expend some effort to maintain the existing interoperability as both Grid infrastructures evolve. In particular, effort will be required to maintain interoperability as LCG migrates from the present RB to the one based on gLite.

In addition to being a technical challenge for job submission, interoperability at the WMS level also poses some accounting challenges. It is to be expected that both LCG and OSG

will require appropriate accounting of the utilisation of their resources by whichever Grid WMS is used to submit to the resources of their respective Grids. We strongly encourage LCG and OSG to sort out these accounting matters in order to avoid administrative hurdles to using one Grid WMS to access both Grids.

As a fallback solution, CMS will implement the CMS workflow management layers described in Section 4.8 such that it can accommodate independent Grid WMS infrastructures on the Grids CMS uses.

4.8 CMS Workflow Management System

In this section we describe the baseline CMS workflow management (CMS WM) system. The CMS WM system manages the large scale data processing and reduction process which is the principal focus of experimental HEP computing. The CMS WM is thus the principal client of the CMS and Grid systems and services described in previous sections in this chapter. While much of the functionality should be provided by components listed above, the CMS WM system must tie together the CMS and non-CMS parts in a way that is useful for CMS physicists.

In this section we will refer generically to the Grid Workload Management Systems as Grid WMS and expect that the CMS WM will provide a single interface for the user with multiple different back-end implementations, one for each Grid (and also one representing local batch queue systems). Data Management services, such as the DBS and the DLS, will be in common, but the actual machinery to submit to one or the other Grid system will be dealt with differently. However as the Grid WMS evolve towards greater interoperability, as described in sec. 4.7.4, we expect that these back-end layers in the CMS WM will get thinner.

In addition there will be slightly different implementations of the CMS WM systems covering data processing systems of varying complexity: analysis systems for the general physicist user and data and Monte Carlo production systems for larger and more complex processing which will be centrally organised by the experiment or by analysis groups. These will be built on top of a common set of CMS WM tools, but will vary primarily due to specialisation for the particular workflows they must implement.

In this section we begin with a basic example distributed workflow to illustrate how the CMS WM system interacts with the middleware and other services. This basic distributed workflow is to a large extent that which will be needed for standard user analysis. In the later subsections we then discuss the more complex production workflows and their additional requirements on the CMS WM system.

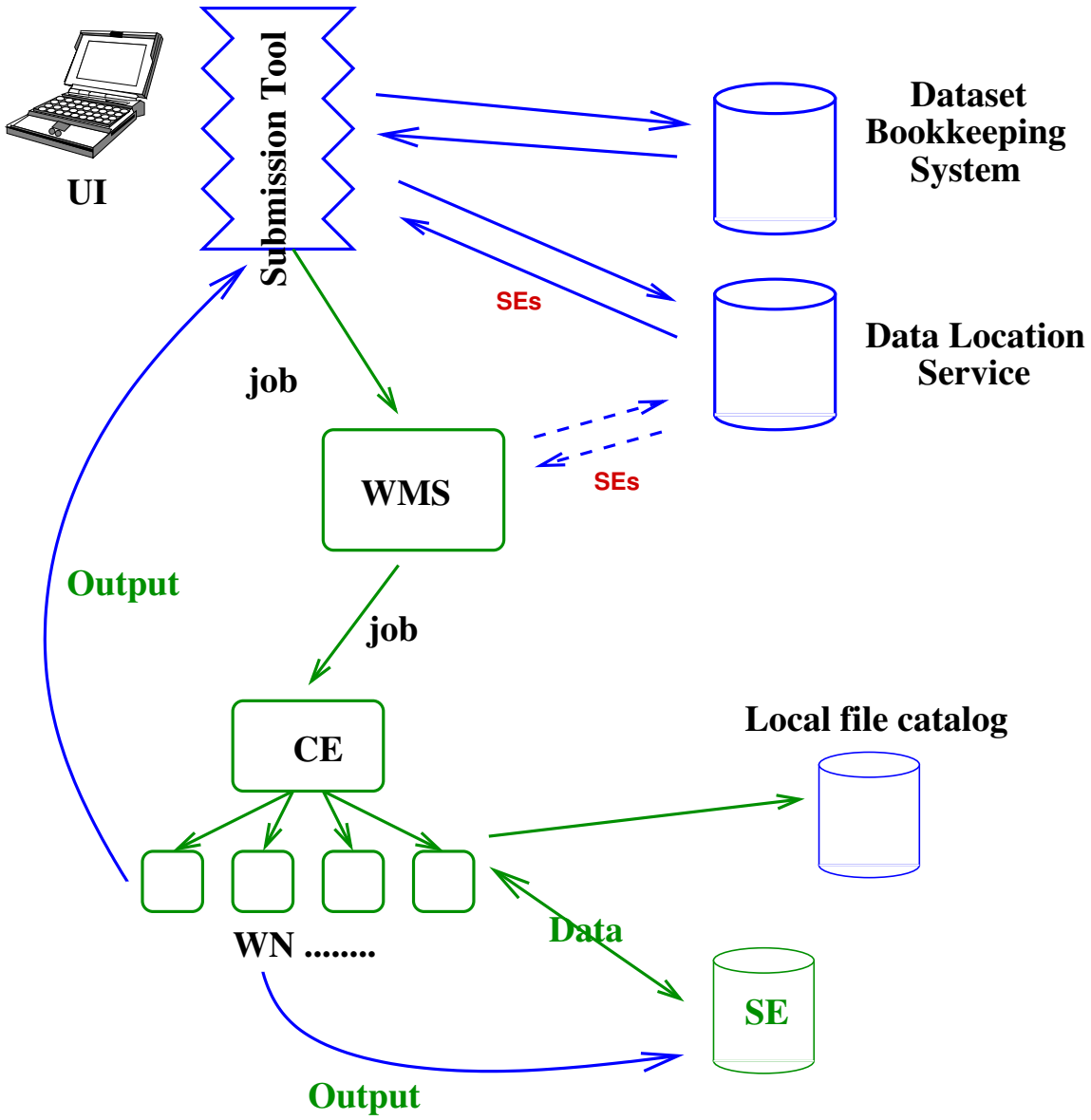


Figure 4.4: Distributed data analysis

4.8.1 Basic Distributed Workflow

We now describe the proposed CMS WM architecture (Fig. 4.4) in the scenario of a single job submission. As described in sec. 4.7 we will largely generalise over the underlying Grid WMS (and also local queue systems) and focus instead on the high-level picture of how the CMS and Grid services are used to implement the workflow.

In what follows we use the following definitions:

- **Task** - This corresponds to the high-level objective of a user such as "run my analysis application over the dataset such-and-such".

- **Job** - The traditional queue system concept, corresponding to a single instance of an application started on a worker node with a specific configuration and output.

It is the goal of the WM system to transform the task formulated by the user into a set of jobs to be run on the distributed computing system and to manage the details of running those jobs on behalf of the user.

The workflow used for the example is essentially that of typical user analysis, but also should cover some simple "production" workflows such as AOD production and analysis group skimming.

The basic steps of the simple workflow and their use of CMS and Grid systems and services will now be described. For simplicity we omit the description of the packaging of the user executables, libraries and the input sandbox. We describe each step in turn and also indicate both where it is taking place (UI, WMS, WN) and which systems/services it uses.

1. **Task formulation:** The user will decide which application they wish to run, which basic application configuration they want (exclusive of data inputs) and decide the general requirements on the input dataset to use.

2. **Data Discovery:** A user (or a program such as CRAB) performs a query to the Dataset Bookkeeping System to find which data is to be accessed. In this query, the user may choose from among existing named datasets (*e.g.* "the latest reconstruction of a particular primary dataset") as well as specify his or her own requirements on the data attributes (*e.g.* software version used for production, calibration tag used, data quality flags, parameter set used, ...).

   The result of the query is a list of event collections, grouped by the underlying file blocks to which the data corresponds. Note that at this stage there is no need to have information about exact data location. It is enough to know that the data exist somewhere, perhaps even in more than in one place. Also there is no need to know about the physical structure of the event collections, this will only be needed further down in the workflow.

   It may be that the user already knows the dataset that he want to access, *e.g.* if he is re–accessing to same data with updated software. In this case, there is no need to perform another query unless the user want to know whether new data has appeared in the system and added to the Dataset.

   The *Data Discovery* step will happen on the User Interface (UI) and involves the DBS.

3. **Job splitting:** At this stage, the WM tool can decide how (and if) to split the complete set of event collections among several jobs, each of which will access a subset of the event collections in the selected dataset. The splitting mechanism will

take care to configure each job with the proper subset of data blocks and event collections.

The user must provide the WM tools with the criteria by which the job splitting will take place (e.g. maximum number of events per job, maximum job run time, etc.).

Job splitting by the CMS WM system takes place (in the CMS baseline) on the UI.

4. **Job configuration:** the WM system will create job configurations for every job which is to be submitted. There are in fact two levels of job configuration: the first for the CMS software framework, the second for the Grid WMS.

   The framework configuration will in general contain only site-independent information (as described in sec. 4.7.1.5). In practice this means that each job will be configured to run on a specific set of event collections chosen as part of the job splitting.

   The configuration file intended for the WMS (e.g. a JDL file) will contain any and all necessary information needed by the WMS to make a decision as to where to dispatch the job. In the CMS baseline, the WM tools will use the File Blocks associated to the event collections assigned to each particular job (obtained during Data Discovery) and contact the DLS and determine which Storage Elements (SE's) have those blocks. The resulting list of SE's will be included in the job WMS-configuration file (e.g as Input Data for an EGEE grid job submission).

   In a beyond-the-baseline scenario, the list of File Blocks for each job may be included directly in the JDL, effectively deferring the DLS lookup to the Job Scheduling step by the Grid WMS (see below).

   The Parameter Set database will be used to extract any and all the parameters needed by the application: these parameters are already defined into the PS database, as a prerequisite for running the application. These parameters will be packed in a site–independent way, and shipped together with the job typically in a file–based fashion. The methods to insert and extract the needed parameters set from the PS database is provided by the application framework.

   If the output data is large enough that it needs to be handled by something other than the output sandbox, the user must also specify at this point the expected final destination of the output data.

   Job configuration takes place on the UI and uses (in the baseline scenario) the DLS and (probably) the Parameter Set database.

5. **Job submission:** After the last step two config files exist for every job in the task: one for the WMS and one for the of the application framework.

   As a fallback solution, the decision about which WMS will be used will be taken at UI level, at submission time. This can be seen as a high level brokering, and will be performed in a very simplistic way, without trying to obtain complete optimisation.

At submission time, the submission tool will have information about data location, and so can apply a simple logic: if data is accessible only via one Grid system, the corresponding WMS will be used. If both Grids provide access to data, a decision based on the UI location can be acceptable.

When the interoperability between the different WMS will be achieved then the WMS chosen will not prevent submission of the jobs to anywhere, so the data will be accessible using either system. As before, a decision based on UI location, or eventually, by user choice will be taken.

The CMS WM tools will submit the jobs to the Grid WMS, as a "job cluster" if necessary for performance or control reasons, and will interact with the job bookkeeping system to allow the tracking of the submitted jobs.

Job submission takes place on the UI (using the tools appropriate for the Grid WMS) and uses the job bookkeeping system and the Grid WMS.

6. **Job scheduling:** The Grid WMS is responsible for scheduling the jobs to run on specific CE's and dispatching them to the CE.

   In the baseline scenario, the WMS will schedule and submit the jobs to a suitable CE using the list of SE's in the WMS configuration file. In the non-baseline scenario the WMS may contact the DLS itself to determine the list of SE's from the list of File Blocks, and thus eventually could perhaps even trigger data replication in some cases to satisfy the needs of the ensemble of pending jobs.

   In both cases, the choice of a CE can exploit not only the data availability but also the data access cost estimate provided by DLS. In the first case (DLS accessed from UI) this cost should be passed together with the configuration file, in the latter case this would be accessed directly from WMS.

   Job scheduling is done by the Grid WMS.

7. **Job run-time:** The jobs arrive on the CE with an application configuration which is still site-independent. As described in sec. 4.7.1.5 it should be possible for the job to determine at the site the locations of necessary site-local services (local file replica catalogue, CMS software installation on the CE, access to CMS conditions, etc.). This will be done by some combination of the job wrapper and (perhaps) the (framework) application itself. Together they must take care to recreate a working environment identical to the original one used by the user to test their application on the UI.

   For example, as the data information in the configuration file of the framework application is site independent, it must be mapped into physical files by the Local File Catalogue. The catalogue is POOL compliant so it can be used directly by the application.

   In addition, if real-time monitoring is being used the job or job wrapper may need to contact the job real-time monitoring database.

Job run-time takes place on a Worker Node of a specific Computing Element and uses the Local File Catalogue, the site storage and data access systems, local CMS software installation, CMS conditions data and (perhaps) the job real-time monitoring system.

8. **Job completion:** Once the job completes, it must store its output someplace. For very small outputs, they may just be returned to the submitter as part of the output sandbox. For larger outputs, the user will have choices depending on the desired destination of the job output. Either the output can be stored on the local SE (for subsequent retrieval by the user from the UI) or the output will be handed off to an agent of the data transfer system for transfer to some other SE.

   In any case, handling of the output data will be asynchronous with respect to the job finishing. The job's only obligation is to either successfully store it to the local SE or pass it to the data transfer agent.

   It is assumed that the Grid WMS will handle making available to the user the output sandbox, log files, etc.

   Job completion takes place on the WN and uses the SE and/or an agent of the data transfer system.

9. **Task monitoring:** While steps 6-8 are in progress, the production or analysis user will monitor the progress of the jobs constituting their task by using the job bookkeeping and monitoring system.

   Task monitoring takes place on a suitable UI and will use the job bookkeeping and monitoring system. The job bookkeeping and monitoring system talk to the Grid WMS and (perhaps) talk to agents on the CE for real-time monitoring information.

10. **Task completion:** As individual jobs finish (or after the entire set of jobs in the task has finished) the user will find the resulting output data coalesced to whatever destination was specified during the "job completion" step above. If the user wishes to publish this data, the relevant provenance information must be extracted from the job bookkeeping system, etc. and published to the DBS. The location of the resulting file blocks can then be published in the DLS.

    Task completion takes place on the UI and uses the job bookkeeping system, the DBS and the DLS.

These pieces thus constitute a basic workflow using the CMS and Grid systems and services described in the earlier sections of this chapter. The CMS WM tools are responsible for orchestrating the interactions with all necessary systems and services to accomplish the specified task.

4.8.2 Prompt Reconstruction

The Prompt Reconstruction (PR) system will be responsible for reconstructing the data arriving from the High Level Trigger (HLT) in quasi-realtime. This first pass reconstruction of the data serves many purposes:

- It provides a first version of the RECO data for further detector, offline calibrations and data quality studies
- The RAW data will be split from its organisation as "online streams" into a file packaging based on the "primary dataset" classification.
- Additional monitoring of data quality beyond that which happens in the HLT
- It provides the FEVT (RAW + RECO) version of the data. A copy of this will be sent to at least one Tier-1 site so that there is a 2nd copy of the RAW data and access to the RECO data outside of the CERN T0.

The PR system is in some ways simpler than the example workflow above. The entire process will happen within the Tier-0 and thus the "Grid WMS" will be replaced by a local batch queue system. The use of local Tier-0 resources also insures that debugging problems will be simpler than in the full distributed computing environment.

The PR workflow is however more complicated in many ways. CMS has yet to specify whether the splitting of the RAW data from the "online streams" into "primary datasets' will happen in a dedicated step just before the reconstruction or as part of the reconstruction itself. In addition, it is possible that the "online stream" data arriving from the HLT may have data taken at approximately the same time clustered into more than one file due to the fact that the HLT will have a "sub-farm" substructure. There is some desire (e.g. for luminosity tracking reasons, access to calibrations, etc.) to reorganise the events such that data taken at around the same is grouped together.

In addition, rather than working on a full complete dataset defined in advance, data will arrive (probably grouped by "run") as it is taken by the CMS detector. It is likely that a dedicated mechanism will be needed in order to recognise that new runs are available. Regardless of whether the online streams or data organised as in primary datasets are used as input to the reconstruction itself, some amount of policy may need to be applied in choosing which streams/datasets to process first. In addition, the constraint that additional latency may be required due to the Prompt Calibration step described below (or perhaps even to wait for calibrations being done "by hand") will quite possibly make the management somewhat complex.

Independent of which solution is taken to manage this problem, a complicated extra step of splitting and/or merging the data will exist in the workflow, only the final output of which must be published to the data management system.

4.8.3 Prompt Calibration

The most critical calibration step is the one needed to reach the accuracy required by the Prompt Reconstruction. If calibrations performed in the online system will not be able to provide such an accuracy, a dedicated "prompt calibration" (PC) step before Prompt Reconstruction will be required. The PC processing will take place in the CMS-CAF.

The PC system will consume the output of the event filter physics or dedicated calibration streams and apply the calibration algorithms in an automated fashion. Such a step is potentially very resource hungry (it may require several reconstruction iterations). Even more importantly, it can easily dominate the overall latency (PC plus PR) for making the event-data available for subsequent analysis and detector studies.

The PC system should in general be very similar to the PR system described above in terms of inputs, however the final "output" will not be event data to be handled by the data management system but instead calibrations to be stored in the conditions database. If intermediate event data formats (e.g. smaller custom calibration ntuples) are used as part of the PC workflow, it may be desirable to store those, but this would just be done using the standard publishing to the data management system.

Offline Prompt Calibration will likely be the most critical activity in the whole analysis workflow. It is therefore essential to optimise the data flow required by this activity, hence we include a few notes on this topic.

We anticipate that calibration and detector-monitoring procedures will mainly use event-data from the express physics stream, from dedicated calibration streams, and eventually from a diagnostic and debugging stream (problematic events). Dedicated streams from the online should help insure low-latency and reduce the need for extra data handling or filtering steps before PC.

The readout of calibration triggers not directly used for physics may be processed differently even at the online level. Partial detector readout (selected sub-detectors only, where regions of interest around lepton candidates are extracted from the whole detector) can be exploited to maintain a manageable data-rate even with calibration trigger rates in the kHz range.

4.8.4 Data Re-reconstruction

It is in general expected that there will be significant improvements to calibrations and software after the initial reconstruction of the data. CMS will thus decide periodically to rerun the reconstruction over the RAW data using the latest software algorithms and calibrations.

For data re-reconstruction, the jobs must run over the raw data. As described earlier there will be one copy of this at the T0 and one copy spread around the T1 sites. When

the T0 is not in use for other things (i.e. no data is being taken and there is no Heavy Ion data to reconstruct), re-reconstruction will be a simpler version of the PR step described above. No online stream to primary dataset splitting will be necessary as the RAW data organised by primary dataset can simply be retrieved from the T0 MSS instead of arriving from the HLT.

In the case where the RAW data copies distributed around the T1 sites are used to do the re-reconstruction, the process should be very similar to the example workflow. The resulting new version of the RECO data will simply be stored at the same site where the RAW data is located.

The CMS baseline does not include dynamic movement of data in response to job submission, but does of course foresee that data be moved intentionally in order to satisfy needs known in advance. The case of data re-reconstruction using the RAW data in the T1's is an example of this. As any given stream may be located in a single custodial T1 site, the need to re-reconstruct a given stream at high priority may exceed the computational resources at a given site even if the data is located there. In this case a rather straightforward optimisation can be done by replicating some of the data blocks to another site temporarily. As the granularity of the replication is relatively large, it can be easily done without initially introducing the complexity of dynamic data movement.

4.8.5 Offline Calibration studies

The ultimate detector accuracy and resolution will be achieved by detailed studies and precise calibration procedures that will require high statistics. These studies will use the result of the Prompt Reconstruction and will produce calibration data to be used for subsequent reconstruction passes.

Examples of such calibration procedures include ECAL crystal intercalibration from samples of high P_T electrons, tracker and muon detector alignment strategies using in-situ tracks, and jet energy calibration. Procedures could take weeks to months after the prompt reconstruction.

This is essentially an analysis type processing as in the basic workflow with individual groups working independently to understand all details of the performance and calibration of a sub-detector. But unlike physics analyses, it requires access to reconstruction and sometimes to RAW data: passes over large samples of RAW (and reconstructed) data will have to be centrally scheduled and coordinated.

A first definition of the RECO is now available, and we have started to evaluate calibration tasks and the required data types in order to understand what changes should be made to improve the usability of the reconstructed event-samples. For instance, one can envision adding close but unassociated hits to each track to allow pattern recognition iterations after alignment, without going back to RAW data. Studies performed using test-beam and simulated data have also shown that the use of a specialised "compact" data-type

can substantially reduce the amount of data required by calibration procedures. Thus a first pass over RAW/RECO data to produce a "compact" data format, specialised for calibrations and easily transportable to a Tier-2 or Tier-3 site may be a common use-case for offline calibration studies. Subsequent analysis may then just happen locally within the Tier-2 or Tier-3 site. This is just the basic workflow described above.

4.8.6 Monte Carlo Production

Monte Carlo (MC) production differs from analysis more in emphasis than in nature. Monte Carlo production is larger-scale than any single analysis, has longer chains (i.e. processes more data-tiers with multiple applications), but is more constrained and deterministic, i.e. the parameters and applications do not change as often as they will for a rapidly evolving analysis. With respect to data production there is also less need to worry about changing and or evolving calibrations.

MC Production (like data reconstruction) has turnaround-times which are long compared to those required for a typical analysis, and the outstanding set of MC requests (and thus jobs to run) is always long. Policies and priorities applying to MC Production are likely to be stable on long timescales compared to those of analysis, especially in the early years.

Where MC Production and analysis are more similar is in the need to have robust proof of what the jobs have done and to be able to guarantee to process all input data exactly once. The bookkeeping needs are identical in functionality, if not scale.

None of this implies that MC Production particularly needs a different architecture. At most it needs a different implementation to support it's current activities, an implementation which should converge with that for analysis as analysis matures. Alternatively, it would use the same components as analysis, but in a different way to analysis. Longer term, as policy-management matures, Production would just become a distinct role sitting in one corner of CMS 'policy-space', rather than a separate activity with its own tools and framework.

More specifically, MC production will to some extent map to the basic workflow above, but does have a number of distinguishing characteristics. First, a system for managing user requests for MC samples as well as their priorities is an integral part of MC production (e.g. this is the "Virtual Data Catalogue" portion of the current RefDB [31]). The workflow is also significantly more complex:

- The whole workflow consists of multiple steps (generation, simulation, digitisation, reconstruction) where the intermediate data often needs to be made available to the end-user.
- Dedicated "merge" steps will be necessary to insure that the file sizes are reasonably large and not determined fully by the output possible within a single job. This implies that tracking of transient files which will later be merged.

- MC production is usually considered a task which can be used to keep CPU capacity busy between peaks from analysis jobs (and also a nice candidate to use resources opportunistically). This complicates scheduling and data management since the planning for opportunistic use is clearly not possible.

These complexities mean that, in addition to the dependencies spelled out for the basic analysis workflow, that the MC production has its own MC request tracking system and (probably) some internal catalogues and data location to track transient files before they are merged.

4.8.7 PROOF and interactive analysis

The CMS Event Data Model (EDM) is currently evolving in the direction of an event persistence that allows ROOT to be used directly for interactive analysis of standard CMS event data.

This opens up the possibility to use the Parallel ROOT Facility (PROOF) as an integral part of the CMS distributed computing system. At present, we expect that some Tier-2 centres across the world may provide PROOF as one of the local services they provide for their users. We do not expect PROOF to be fully integrated into the global CMS computing system as part of the baseline functionality and do not expect that it will be required at all Tier-2 sites.

As part of this baseline "site-local" use of PROOF, data would in general be accessed within a given site so use of the Grid WMS, DLS, etc. would not be necessary. Users will still perform "Data Discovery" and "Job Configuration" steps even within the context of ROOT/PROOF. Simple additions (either in collaboration with the ROOT team or in CMS-specific additions) could facilitate (for example) accessing the DBS and avoiding the management of site-specific information themselves in their configuration files.

Beyond the baseline, we foresee that PROOF sessions may become better integrated with the full CMS workflow management in order to use distributed resources beyond those of a single site. For example, a user might request to analyse a given dataset registered with a DBS using PROOF. The basic logic for satisfying such a processing request would be no different from that described in Section 4.8.1 for batch processing. CMS WM would identify a site that supports PROOF and has the desired dataset stored locally. A PROOF master and a set of slaves are started at that site, and the interactive session connects to this PROOF session. The interactive session might steal cycles from CMS batch operations up to some high watermark. Given that the CPU duty cycle for interactive work is generally rather low, the interactive session might have first priority on the CPU as long as this high watermark is not exceeded on average. In a Grid WMS architecture as described in Section 4.7.1 it is in principal possible for CMS to know exactly how long it may utilise a given slot in the site batch system. Adding for example parasitic running of PROOF on top of this architecture is thus in principle straightforward.

In practice, a complete integration of PROOF into CMS workflow management would require significant effort beyond the scope of the baseline architecture. A detailed description of how PROOF might be integrated into CMS workflow management, as well as the underlying Grid WMS is thus beyond the scope of this report.

4.8.8 Interoperability of WM systems

Since we expect that there will be slightly different flavours of CMS WM systems used in environments of differing complexity (data production, MC production, analysis, etc.) it is important to note that these should work together without requiring the consumer of data produced by one of those systems to work with more than one at the same time.

This is done in two ways: first, underlying each of the WM flavours should be some set of common tools on which they are based. Second, as shown in fig. 4.5, the systems will communicate with one another by *publishing* the data they produced to the DBS. Subsequent consumers then discover data from the DBS. Since the publishing step will effectively drop unimportant information (e.g. the fact that a given WM system merged data produced by separate jobs), the DBS acts as the means to communicate between these systems. Similar arguments apply also for the DLS data management component.

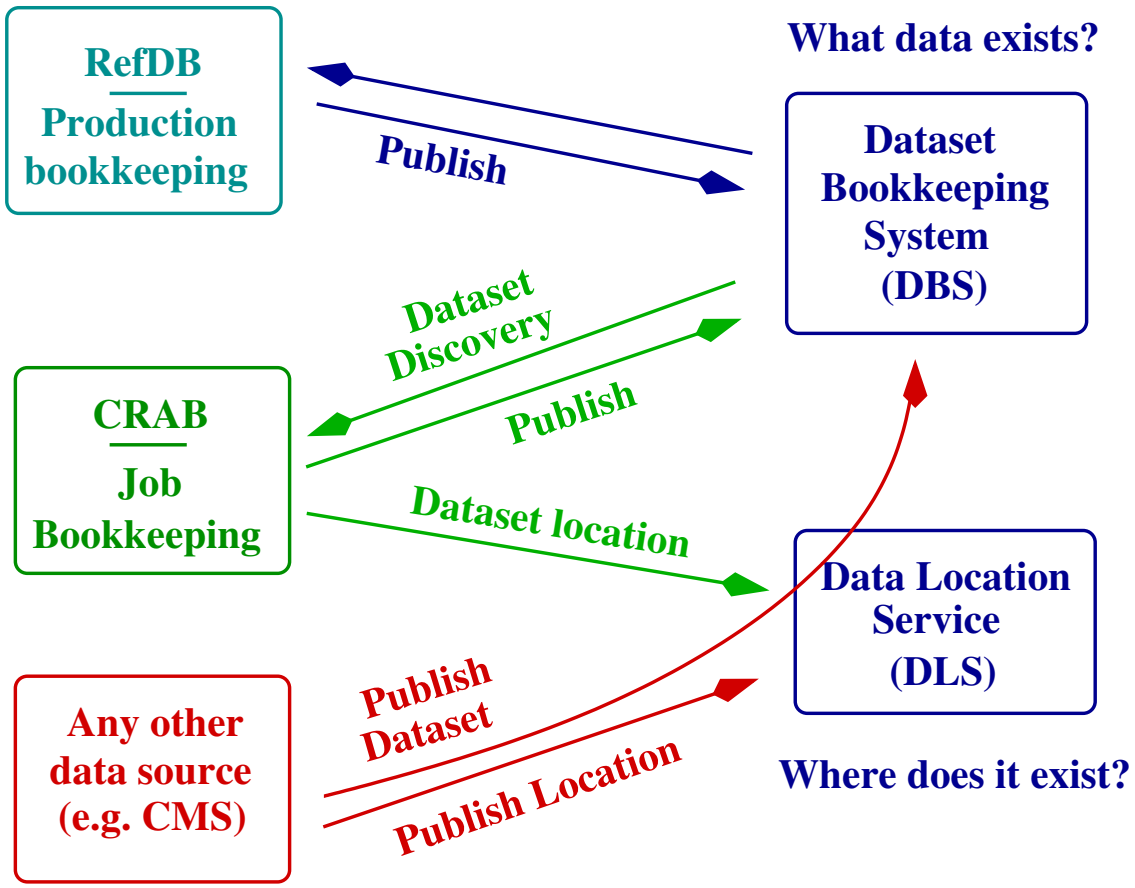


Figure 4.5: Relationship between production/WM systems and Dataset Bookkeeping System

4.9 Status of Components of the Proposed Computing Services

In this section we summarise in a few tables the current status of the design, implementation or deployment of the various components we have identified as part of our Computing Services in this chapter.

Data Management	
RefDB/PubDB	Legacy DM system created in particular for MC production, being refactored into the new DBS/DLS components to cleanly separate DM from MC workflow support
DBS	In very early prototyping and design stage
DLS	Simple, temporary mock-up available to allow prototyping of other pieces, beginning evaluation of Grid components/catalogues to determine if they can be used for this component
Local file catalogues	Simple system exists since some time in the form of simple xml catalogues. Need to move to "site-local" discovery of catalogues and explore options for simplification of catalogue management.
Data Placement and Transfer System	The reference implementation is the current PhEDEx system. It implements the basic functionality required, although evolution is needed to work properly with the new DM system (DBS/DLS, local file catalogues) and to support specific workflows (e.e. such as harvesting/coalescing of data from analysis jobs). Missing also support for small and/or transient systems and managing priorities.
Data Access and Storage Systems	Provided mainly by "external" organisations, although CMS must evaluate the products (Castor, dCache/SRM, DPM, xrootd)

Table 4.1: Status of Data Management Components as of June 2005.

Application and Job System Services	
Parameter Set Management System	Does not exist, still in the definition phase.
Job Bookkeeping and Monitoring System	A reference system with the basic functionality exists (BOSS). The system is however missing support for "lightweight" (e.g. SQLlite) databases, a real-time monitoring framework with independent DB server and support for "tasks" (job groups).
Conditions database	Does not exist, still in the definition phase.

Table 4.2: Status of Application and Job System Services as of June 2005.

Grid Workload Management Systems	
Resource Broker	Working smoothly since some time. Outstanding issues are related to performance and bulk job submission. Waiting for gLite RB to resolve these issues.
Data Location Interface	Implemented and tested in a testbed with interface with the old DM (RefDB/PubDB).
Job Policies and Priorities	The possibility to define these dynamically at VO level does not exist, still in definition phase

Table 4.3: Status of Grid Workload Management Components as of June 2005.

CMS Workflow Management	
Basic Analysis Workflow	A basic system for job splitting and submission exists (CRAB). It currently uses the old DM system from MC production (RefDB and PubDB) and uses a very simple, custom job bookkeeping/monitoring system instead of the standard one.
Prompt Reconstruction	No support currently exists for this workflow.
Prompt Calibration	No support currently exists for this workflow
Data Re-reconstruction	No support currently exists for this workflow
Offline Calibration Studies	No support currently exists for this workflow
Monte Carlo Production	Extensive support for this workflow exists, although geared to the "old EDM". This includes basic data management infrastructure (RefDB/PubDB) which will need to be migrated to the new DM components (DBS/DLS, etc.) and the "new EDM".
PROOF and interactive analysis	No support currently exists for this workflow.

Table 4.4: Status of CMS Workflow Management Components as of June 2005.

Chapter 5

Computing Project Management

The Computing Task is part of the CMS CPT project[1], which contains four main tasks: Analysis-PRS[2], Detector-PRS, Software and Computing. The Analysis and Detector tasks are an intrinsic part of the physics research program; they are not the subject of this TDR.

The CPT project reports to the CMS Management Board and the associated CMS Steering Committee. Collaboration oversight is provided by the CMS Collaboration Board and its subsidiary Computing Institution Board (CIB). Liaison with CMS computing centres is within the purview of the CMS Computing Committee (CCC) whose membership comprises two representatives (a physicist and a computing expert) for each Tier-1 including CERN and representatives of regional groupings of Tier-2 centres.

Computing and Software resources are within the purview of the CPT Resource and Planning Manager who chairs the Computing and Software Finance Board, which includes representatives of Computing and Software resource providers, and reports to the CMS Finance Board.

The Computing Project Management is described in the following sections:

- Computing Project Scope and Responsibilities (Section 5.1),
- Computing Project organisation(Section 5.2),
- Computing Project schedule and milestones (Section 5.3) and
- Computing Project resource needs(Section 5.4).

[1]CPT as the name for the project is a historical term in CMS, where C denotes Computing, P denotes physics analysis / detector software, and T denotes (Higher-Level) Triggers

[2] PRS is a historical term in CMS, that denotes the Physics Reconstruction and Selection groups

5.1 Computing Project Scope and Responsibilities

The Computing Task has responsibility for delivering the computing systems and services for CMS; the main body of this TDR describes the technical scope. These systems will use the computing infrastructures and services being provided by the Worldwide LHC Computing Grid, WLCG. The WLCG is a worldwide collaboration of computing centres and Grid projects that is described in the draft WLCG Computing Grid Collaboration Memorandum of Understanding [13]. The Computing Task is responsible for developing the technical baseline for these services including their interaction to form a coherent and friendly CMS Computing Environment for CMS users. The Computing Task will, in addition, provide and develop an architecture of application specific components and services that interact with the services and resources provided by the regional computing centres and CERN as part of the WLCG.

In addition to providing CMS computing services and interfaces between resources and applications, the Computing Task has responsibility for integrating these services into a coherent and functional infrastructure that supports the CMS workflows (MC production, data analysis and processing, and calibration), and that is usable for individual physicists as well as for production managers.

The Computing Task is also responsible for the operation and execution of many of the CMS workflows, and to develop and execute an operations plan that supports all CMS computing users in their work using the CMS computing systems. This operations activity in many cases will connect CMS users with the operations support systems of the WLCG computing providers and services.

In addition, the Computing Task is responsible, working with the operations and IT support teams at regional centres, including CERN, for making sure that facilities and infrastructure services are being provided to CMS in an efficient and appropriate manner in accordance with CMS scientific priorities. In particular, CMS will have both a CMS Tier-0 and a CMS CAF Facilities Coordinator at CERN. CMS will also have personnel at CMS Tier-1 centres to liaise with operations staff and IT provider teams.

In addition to the resources for providing CPU, data storage, and access services, the distributed CMS computing systems at CERN and offsite includes:

- infrastructure for software development, installation and distributed deployment, and for databases, information and documentation services;
- building, deploying, and operating systems for general and user-specific data productions;
- safe storage and distribution of CMS data; and
- workflow support and large-scale data processing for Monte-Carlo simulation, reconstruction, and analysis activities.

The Computing Task is responsible for the further development of the CMS Computing

Model and the technical baseline and its validation, to arrive at an initial end-to-end system for the start of data taking.

The Software technical design will be described in the Physics TDR, and is not the subject of this TDR. For clarity, the software task includes the responsibility for delivering: (1) the core application software, (2) the software for physics and detector simulation, reconstruction, calibration and physics analysis, (3) the software to implement Higher Level Triggers and the associated algorithms, and (4) the software to assure the quality and integrity of CMS data.

5.2 Computing Project Organisation

The organisation chart within "Computing" is shown in Figure 5.1. The Computing Coordinators provide overall leadership to the Computing Task. They are members of the CPT Management Board and represent the experiment in all matters of computing to the outside of CMS. They are responsible for developing the project plan and schedule, for allocating resources working with the contributing institutions, and for executing the project plan.

The Computing Task itself is being organised into four program areas, the Technical Program, the Integration Program, the Operations Program, and the Facilities and Infrastructure Program. Coordination of each of these Level-2 program areas is delegated to their respective program coordinators. Together with the Computing Coordinators, the coordinators for the Technical, Integration, Operations, and Facility Programs form the Computing Management Team.

A more detailed management plan including the work plan, schedule, milestones will be developed and maintained as part of the CPT organisation, as the planning for CPT evolves.

Work is performed in a set of Computing sub-projects from which all the four program areas draw effort. These sub-projects are formed to deliver specific services, products, documents, or other deliverables, according to the overall plan. They are typically organised with a project lead reporting to one of the Computing program areas (typically to the Technical Program Coordinators), with each of the Technical, Integration, Operation and Facilities Programs as stakeholders in the sub-project deliverables and schedules.

People working on the Computing Task are being assigned to these projects. CMS strives for each sub-project to identify one or more lead institutions to be responsible for each of the projects and to provide effort and continuity.

Effort from these projects is being provided typically to the technical program, as well as the integration and the operations program. The project schedules are being negotiated between the project lead(s) and the Computing Management Team. Specifically, the needs

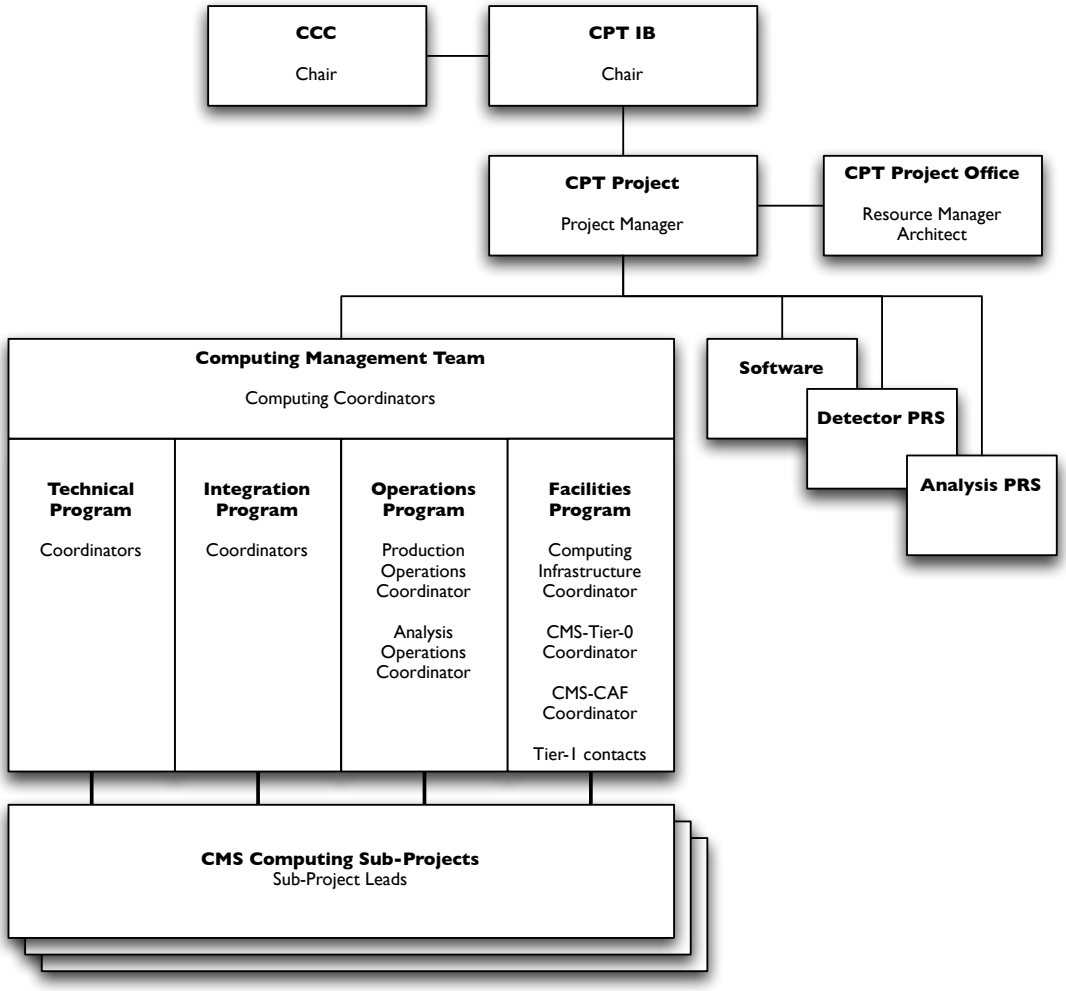


Figure 5.1: Organisation Chart of the Computing Task in the CPT Project

for all project areas will be considered when planning for any of the specific sub-projects. Thus, while the Technical Program coordinators are the stakeholders for the technical ("Functionality") milestones of a given project, and in many cases are the coordinators the sub-project lead reports to, the Integration Program coordinators are the stakeholders for delivering required integration effort, and the Operations Program coordinators for the operation of the deliverables of the project.

The Computing Task overall is driven by the CPT Level-1 and the Computing Integration Level-2 milestones. The Computing Management Team negotiates the exact scope of each of the Level-2 milestones, in particular as they relate to the work program of the sub-projects. The scope of each milestone thus propagates to the scope of the sub-project, and their detailed Level-3 milestones.

5.2.1 Technical Program

The Technical Program is coordinated by the Technical Program Coordinators who are working with the Computing sub-projects as described above. The Technical Program Coordinators' main responsibilities are developing and maintaining the CMS Computing Technical Baseline, working with the Computing Management Team, and coordinating the set of sub-projects to deliver well-defined components.

CMS Computing is running a well-defined set of sub-projects as part of the Technical Program to address specific deliverable areas:

- job configuration and scheduling services, including policies, prioritisation of workloads, and ensuring scalability of job scheduling;
- dataset placement and data transfer services,
- dataset bookkeeping services, including physics metadata services, the data quality management infrastructure, luminosity data services, and other environmental information pertaining to datasets;
- instrumentation services and user interfaces, including a CMS dashboard, monitoring, job tracking, also configuration monitoring and validation;
- CMS storage dataset access services, e.g., SRM, Castor, POOL/PubDB++/local file catalogs, including CMS support for site storage services, access control, and data management interfaces;
- CMS workflows support, for the data taking and processing, MC production, and calibration workflows. This also includes configuration and provenance services, software installation support, configuration control, software packaging, and interactive analysis tools support;
- CMS VO services, including CMS support for VO management, privilege, accounting, and security.

Technical functionality milestones are being set for each of the sub-projects, that each

have a defined scope and lifetime, together with a set of deliverables. The technical plan follows the computing baseline, and the overall schedule is being negotiated in the Computing Management Team working with the Integration task as part of the CMS computing integration plan.

5.2.2 Integration Program

The Integration Program is coordinated by the Integration Program Coordinators. We expect to form an integrations team, but most of the component integration effort will actually be delivered by the technical sub-projects. The Integration Program will require a specific computing infrastructure, the Integration Computing Environment. This Integration environment should be provided to CMS as part of the WLCG and in collaboration with the Grid projects and the LCG project.

A CMS Computing Integration Plan will be developed that describes the integration program of work and schedule. The set of integration milestones provide the CMS Computing Level-2 milestones, as described in section 5.3.

The responsibilities of the Integration Program Coordinators include

- developing and maintaining the CMS Computing Integration Plan, working with the Computing Management Team;
- preparing for and running of series of Computing Integration Milestones, data challenges, and service challenges;
- taking responsibility for validation, releases, and deployment while working with technical, operations, and facilities programs;
- providing workflow integration for production, dataset publishing, distributed analysis, data taking and processing;
- providing component integration, working with the technical program on integrating the CMS computing systems and components, into a coherent and functional computing environment;
- releasing and delivering the integrated computing environment of CMS computing services and components into the CMS production Grid environment, working with the Technical Program, the Operations Program and the WLCG; and
- liaising on a very practical level with CERN-IT,the CMS regional centres, the LCG and Grid projects, their technology, fabric and facility providers, and with the Grid consortia: EGEE, OSG, and NorduGrid.

5.2.3 Operations Program

The Operations Program is coordinated by the Operations Coordinators. There is an Operations Coordinator for MC Production Running and a Coordinator for CMS Computing User Support.

Much of the operations effort will need to be delivered by regional centres and by the Grid operations centres and services. The CMS Computing technical sub-projects have responsibilities to the Operations Program, in that functionalities to operate components as part of the WLCG and the CMS Computing environment will be provided by these projects, working with the Technical Program and the Integration Program for delivery.

The Operations Program has milestones related to developing an Operations Model and Operations Plan and keeping it updated, and milestones related to delivery of functionality to run operations. These Operations Level-3 milestones will be specified in the Operations Plan and are each associated to specific Computing Level-2 milestones.

The responsibilities of the Operations Program coordinators include:

- developing and maintaining the CMS Computing operations model and plan, working with the Computing Management Team,
- MC production operations,
- database system operations,
- calibration workflow support,
- data-taking operations and data validation; and
- user support.

5.2.4 Facilities Program

CMS Computing recognises the importance of the computing facilities providing the required resources to the experiment. The Facilities Program area is to coordinate with the resource providers and the operations teams at the regional centres, to ensure the provision of these resources and facilities according to CMS needs and policies.

Facilities are being provided by the IT organisations at the sites, including CERN. As part of the WLCG overall operations is organised through the Grid operations activities, there are important coordination responsibilities between these organisations and the CMS experiment Computing Task. The Facility coordinators provide this coordination.

At CERN there are the particular tasks of working with the CERN IT department and the LCG project on delivering the CMS Tier-0 services for CMS data taking and data processing, and on making the CMS CAF available and operational for CMS users. The CMS Tier-0 facilities coordinator and the CMS-CAF coordinator are functions that will be instrumental to this coordination.

Another important function is the provision of core computing infrastructure and services. This covers services that are a shared need and responsibility of the full CMS Collaboration, such as the support for the CMS common computing and software environment, software process services, and core support for production operations. These services are coordinated by the CMS Common Services and Infrastructure Coordinator.

In preparation for data taking, CMS will need to keep close liaison and coordination with its Tier-1 centres that provide the bulk of computing resources required for data processing and data analysis. Each Tier-1 centre should provide a CMS Tier-1 contact person.

Together, these form the CMS Computing Facilities program, thus consisting of

- CMS Tier-0 coordinator
- CMS CAF coordinator
- CMS common services and infrastructure coordinator
- CMS Tier-1 technical contacts

CMS Computing is an important stakeholder in the program of work for the facilities. The facilities work plan will be coordinated with the Computing Management Team, and a Facilities Plan will be worked out with specific facilities related Level-3 milestones, related to functionalities, scalability, and performance metrics of provided facility services. These facilities milestones will be associated with the Computing Level-2 milestones.

5.3 Computing Project Schedule and Milestones

In this section we describe the general approach to the Computing scheduling, outline the main project phases, and specify the major CPT milestones (Level-1) and the subsidiary Computing milestones (Level-2). These phases and Level-1 milestones drive the Computing schedule, and the Computing Level-2 (Integration) Milestones deliver to these CPT Level-1 milestones.

CMS takes an iterative approach by developing the CMS computing environment with regular "challenges", described below, that test and exercise the state of the computing systems. The Level-2 milestones are aligned to these challenges through the use of regular integration milestones which mark the CMS computing systems readiness for these challenges.

This approach focuses the Technical Program, the Integration Program, and the Facilities and Infrastructure Program to deliver a fully functional and well integrated and tested system for the start of data taking, At the same time it will provide a functional system for each of the challenges. After achieving each of the goals related to the integration milestones, the Integration Computing environment will be moved into production and be used for MC data production, user analysis running and dataset serving for physics prepa-

ration, and testing of calibration procedures as part of and supported by the Operations Program.

This approach also allows the Computing Task to regularly assess readiness of the system and its individual components. In addition, Computing will always provide a working system, while allowing progress to be made on the functionality, scalability, and robustness within the Integration Computing environment. Upgraded versions of the Computing systems will be regularly supplied to the Operations Program in the production Grid environment. Each Computing Integration Milestone thus serves as a check-point for the technical components, operations components, and facility components that get integrated into the next version of the CMS computing environment. That version is moved into production use by the Operations Team.

5.3.1 CPT Project Phases

The major phases of the CPT project are as follows.

Computing support for the Physics TDR (up to Spring 2006)

CPT is responsible in this phase for providing the core software frameworks, services (e.g., graphics), and a development environment for use by the PRS groups for their simulation, reconstruction, and analysis activities. The main task of CPT in this phase is the provisioning of a complete, though not necessarily final, baseline set of software tools as required for the preparation of the Physics TDR. This includes the core software and systems required to operate a coherent and robust distributed computing system. A worldwide system of computing centres is needed to support large-scale productions and subsequent analysis for the Physics TDR activities.

Cosmic Challenge (Autumn 2005 - Spring 2006)

This challenge will exercise data taking through the online system as installed for the magnet test/slice test. This will provide a first systems test of the data taking workflow, including the new core software framework and event data model, and its interfaces to the computing systems, in particular data management, data transfers, and experiment non-event data handling.

These will support the reconstruction of data taken from the detector during the cosmic run after the CMS magnet test. The main computing tasks will be to drive the reconstruction workflow, to deliver and deploy a database infrastructure working with the LCG and the Software database task, and to move the cosmic run data to regional centres, making the information about datasets available to physics users throughout CMS.

Service Challenges (2005-2006)

In this phase the computing systems will ramp-up and the computing services which are defined in this TDR will become available, making the system increasingly more functional. These service challenges prepared and executed together with the LCG project, regional CMS projects, and regional Tier-1 and Tier-2 centres, provide synchronisation points between the CMS systems and the WLCG services at CERN and at regional centres.

The Service Challenges will allow for the assessment functionality and performance metrics of computing services and the system as a whole, as it is being developed. MC production and user analysis computing and dataset serving will need to be supported in order to help the rapidly increasing PRS activities in preparation for the Physics TDR. There will be an increasing reliance on LCG services and the Service Challenges aimed at demonstrating that the WLCG services perform as expected. These will use a time-shared set of distributed centres corresponding to approximately 50% of a single experiment's needs at the time of LHC start up.

Computing, Software and Analysis Challenge (2nd half 2006)

The primary goal of the Computing, Software and Analysis Challenge 2006 is to ensure readiness of the software and computing systems for the first real data from CMS. It should contain a few tens of millions of events produced at the CMS Tier-0, split into a number of physics datasets, the local creation of AOD and its distribution to all participating Tier-1 centres, the distribution of FEVT data to Tier-1 centres, the running of physics jobs on AOD, and, at some Tier-1 centres, of calibration jobs. Tier-2 centres should be included in this challenge to demonstrate the ability to extract parts of the AOD to Tier-2 sites for user analysis and calibration processing.

Staged Commissioning of Computing Systems (1 yr before beam–2 yrs after)

After the initial provisioning of the computing systems for the start of data taking, computing resources will continue to ramp up to support running at increasing luminosities, rising to full nominal LHC luminosity. It is expected that through Moore's law, by exchanging out-of-date computing system components at a roughly 3-year cycle, the required performance and resource increases can be obtained at a roughly constant yearly budget.

5.3.2 CPT Milestones (Level 1)

The full list of CPT Level-1 milestones is shown in Table 5.1. Each of these Level-1 milestones has an associated set of Level-2 milestones. By definition, completion of a CPT Level-1 milestone is achieved upon completion of the complete set of the associated Level-2 milestones.

Milestone ID	Description	Due date
CPT-1	Submission of Computing TDR	**Jun 2005**
CPT-2	Baseline Computing / Software systems and Physics procedures for the Cosmic Challenge and Physics TDR	**Dec 2005**
CPT-3	Submission of Physics TDR (Vols I and II)	**Apr 2006**
CPT-4	Computing, Software and Analysis Challenge (CSA 2006) complete	**Sep 2006**
CPT-5	Computing and Software Systems and Physics procedures ready for data taking	**Feb 2007**
CPT-6	Tier 0,1, and 2 computing systems operational (pilot run capacity)	**Jun 2007**
CPT-7	Tier 0,1, and 2 computing systems operational (low luminosity capacity)	**Apr 2008**
CPT-8	Tier 0,1, and 2 computing systems operational (high luminosity capacity)	**Apr 2009**

Table 5.1: Level-1 Milestones of the CPT Project.

5.3.3 Computing Milestones (Level 2)

For Computing, the work is organised along a sequence of integration milestones. These are Level-2 milestones for the overall CPT project, and are related to an overall CMS and CPT goal and externally controlled schedule item. Examples are the Service Challenges, Cosmic challenge, and the start of data taking.

Currently the Computing Level-2 milestones are defined broadly by specifying the main goals and deliverables. As part of the planning for the Computing Task, each of the upcoming milestones and their technical content will be specified in increasing detail and precision as each of the milestones approaches. A milestone goals assessment and lessons-learned document should be provided after each milestone is achieved.

Each of the Computing Integration Milestones integrates a set of Level-3 milestones related to readiness of computing subsystems and components ("Functionality Milestones"),

Operations Milestones, and Facility Milestones. A first set of these more detailed internal Computing milestones is given in the following sections. The Integration Program works to achieve the Computing Integration Milestone goals, together with the Technical Program, the Operations Program and the Facilities Program as major stakeholders.

Operations, Facilities and Technical Programs develop and maintain a set of detailed Functionality Milestones with the Computing Integration Milestones as major drivers. These programs are "Programs of Work" that are broken down into a set of managed subprojects with (wherever possible) well-defined scopes, deliverables, schedules, resources, and start and end dates that become part of the overall planning and Level-3 milestone structure.

The CPT Level-1 and Level-2 milestones are shown in Figure 5.2. A rolling planning strategy is used whereby the future work plan and associated milestones are refined as time progresses. This is clearly seen in the figure where the plan for 2005 is more detailed than subsequent years. The initial list of Computing Integration Milestones addresses the series of CPT Level-1 milestones above, with the goal of providing the integrated computing systems and environments for these challenges, up to the start of data taking. These will be further defined and solidified in the Computing Integration Plan, in coordination among Integration, Operation, Facilities, and Technical programs. The technical content of each of the milestones will then be defined in sufficient detail as to allow coordination between the different program areas and the different sub-projects.

CPT-202/C: "Initial integration of baseline computing components" (Jul 2005)

The main goal of this milestone is the initial assessment of baseline components described in the C-TDR:

- assessment and definition of project plans for components that make up the computing technical baseline, including prototype version of data management components, including DBS, DLS, production version of MCPS, PhEDEx, refactored RefDB
- assessment of VO services, information services, monitoring services, security and access control, accounting
- assessment of job scheduling performance and scalability

The integration goal of this milestone is the initial integration of the SC3 components, including legacy components, into a system functionally equivalent to the existing workflow for MC production and data transfer services.

Milestone deliverables include the following: project plans for the components forming the computing technical baseline (this TDR), an initial integration program plan, an initial operations model and operations plan, MC production plan, and an initial CMS facilities plan.

Year	L1 Parent milestone	Date (version 34.2)	Milestone title	Level	ID	Responsible
	CPT-1	Aug-04	DC04 (5%) data challenge complete	2	CPT-101 / C	Computing
2005		Jan-05	Computing Model paper complete (1st draft Computing TDR)	2	CPT-102 / C	Computing
		Jun-05	**Submission of Computing TDR**	**1**	**CPT-1**	**CPT**
	CPT-2	Jul-05	First version re-engineered Event Data Model / Framework and low-level detector raw data	2	CPT-201 / DS	Detector & Software
		Jul-05	Initial integration of baseline computing components	2	CPT-202 / C	Computing
		Sep-05	First version of event processing applications for Cosmic Challenge	2	CPT-203 / DS	Detector & Software
		Sep-05	Computing systems ready for Service Challenge SC3	2	CPT-204 / C	Computing
		Oct-05	First version simulation application using re-engineered Event Data Model / Framework	2	CPT-205 / S	Software
		Oct-05	First version calibration / alignment software infrastructure using re-engineered Event Data Model / Framework	2	CPT-206 / DS	Detector & Software
		Oct-05	First specification of procedures for CMS calibration and alignment	2	CPT-207 / DS	Detector & Software
		Oct-05	First version high-level physics algorithms / objects using re-engineered Event Data Model / Framework	2	CPT-208 / DS	Detector & Software
		Oct-05	Detector synchronisation procedures complete	2	CPT-209 / D	Detector
		Nov-05	Demonstration of an analysis application for the Physics TDR using the re-engineered Event Data Model / Framework	2	CPT-210 / ADS	Analysis & Detector & Software
		Dec-05	Detector procedures and Software ready for Cosmic Challenge	2	CPT-211 / DS	Detector & Software
		Dec-05	Computing systems ready for Cosmic Challenge	2	CPT-212 / C	Computing
		Dec-05	**Baseline Computing / Software Systems & Physics Procedures for Cosmic Challenge & Physics TDR**	**1**	**CPT-2**	**CPT**
	CPT-3	Nov-05	Submission of the common TOTEM/CMS LOI on diffraction and forward physics	2	CPT-301 / AD	Analysis & Detector
		Dec-05	Physics TDR Volume I complete	2	CPT-302 / DS	Detector & Software
2006		Apr-06	Physics TDR Volume II complete	2	CPT-303 / ADS	Analysis & Detector & Software
		Apr-06	**Submission of Physics TDR (Volumes I and II)**	**1**	**CPT-3**	**CPT**
	CPT-4	Jan-06	Simulation and digitization software ready for CSA-2006	2	CPT-401 / S	Software
		Mar-06	Computing systems ready for Service Challenge SC4	2	CPT-402 / C	Computing
		Mar-06	Detector and Physics Reconstruction ready for CSA-2006	2	CPT-403 / DS	Detector & Software
		Jun-06	Computing systems at Tier-0, 1, 2 centres ready for CSA-2006	2	CPT-404 / C	Computing
		Jun-06	Calibration, alignment, analysis and visualisation ready for CSA-2006	2	CPT-405 / ADS	Analysis & Detector & Software
		Sep-06	**Computing, Software, and Analysis Challenge (CSA-2006) complete**	**1**	**CPT-4**	**CPT**
	CPT-5	Oct-06	Pre-production software system ready	2	CPT-501 / S	Software
		Oct-06	Computing systems re-visited based on CSA-2006 lessons-learned	2	CPT-502 / C	Computing
		Dec-06	Demonstrate performance of HLT/offline reconstruction, calibration, alignment, visualisation	2	CPT-503 / DS	Detector & Software
		Dec-06	Integration of Computing Systems at Tier-0, 1 and 2 centres	2	CPT-504 / C	Computing
2007		Feb-07	Software complete: HLT, reconstruction, simulation (fast and full), calibration, alignment, visualisation, analysis	2	CPT-505 / ADS	Analysis & Detector & Software
		Feb-07	**Computing and Software Systems and Physics Procedures ready for data-taking**	**1**	**CPT-5**	**CPT**
	CPT-6	Feb-07	Tier-0 centre and CERN Analysis Facility ready for pilot run	2	CPT-601 / C	Computing
		Apr-07	Tier-1 and 2 centres ready for pilot run	2	CPT-602 / C	Computing
		Apr-07	HLT/offline software systems ready for pilot run	2	CPT-603 / S	Software
		Jun-07	**Tier 0, 1, and 2 Computing Systems Operational (pilot run capacity)**	**1**	**CPT-6**	**CPT**
2008	**CPT-7**	**Apr-08**	**Tier 0, 1, and 2 Computing Systems Operational (low luminosity capacity)**	**1**	**CPT-7**	**CPT**
2009	**CPT-8**	**Apr-09**	**Tier 0, 1, and 2 Computing Systems Operational (high luminosity capacity)**	**1**	**CPT-8**	**CPT**

CPT_v34-2-revision-16 Updated: Lucas Taylor 16-June-2005

Figure 5.2: High-level milestones for the CPT project, Version 34.2.

CPT-204/C: "Computing systems ready for Service Challenge SC3" (Sep 2005)

This is the first integration milestone to be scoped out in detail in the integration program plan. The integration goal of this milestone is to get the CMS computing services ready for SC3, in particular:

- PhEDEx data placement and date transfer systems ready for SC3 workflow and integrated with regional centre storage services and file transfer services
- an initial CMS system dashboard
- a prototypical support of data management components and re-factored RefDB for the MC production workflow
- initial look at gLite and pull-model production scheduling

The technical program deliverables include items such as: the data management components supporting the production workflow; a first version of the re-factored RefDB; and an initial CMS dashboard, displaying monitoring information pertaining to the state of the CMS Computing environment. The integration, operations, and facilities deliverables include: the test of the CMS SC3 environment, according to integration plan metric; the CMS SC3 run plan; the initial operations plan for supporting MC production; and the conceptual design of CMS Tier-0 and CMS CAF. The milestone has external dependencies including: status of SC3 testbed preparations, LCG deployments, for example gLite.

CPT-212/C: "Computing systems ready for Cosmic Challenge" (Dec 2005)

The goal of this milestone is to get the computing systems ready to support the Cosmic Challenge workflow. It includes:

- EDM metadata handling interface to the datamanagement components
- re-factord RefDB supports the new framework, is interfaced to DBS, MCPS
- initial database infrastructure and workflow established and deployed
- CMS-CAF supports a significant amount of users for analysis running and user space
- operations supports physics users

The deliverables include: the cosmic challenge computing run plan; initial operations support systems, for example a ticket system. The milestone depends on a number of external, including: software releases of new framework; DM components integrated with framework; calibrations and conditions database deployed; and the availability of post-SC3 LCG environment.

CPT-402/C: "Computing systems ready for Service Challenge SC4" (Mar 2006)

The goal of this milestone is to get the computing systems ready for SC4 tests. This includes: the computing systems supporting a framework for individual users and the establishment of CMS site facility plans and operations plan.

The primary deliverables include: user job configuration system for new framework; job scheduling system at SC4 scale; and analysis environment, with functionalities yet to be listed. Secondary deliverables include: the establishment of the CMS Tier-1 managers; a list of CMS Tier-2s online; data management tools allow CMS to manage data across Tier-1s and Tier-2s; and the initial configuration and resource monitoring and job tracking.

CPT-404/C: "Computing systems at Tier-0, 1, 2 centres ready for CSA 2006" (Jun 2006)

The goal of this milestone is to prepare the computing systems, with tests in July, for the Computing, Software and Analysis Challenge (CSA 2006). This includes establishing operational CMS Tier-0 and CMS-CAF systems (20% capacity) and selected Tier-1/2 centres, likely those we will initially rely on for 2007.

The deliverables include: an operational CMS Tier-0 (20% capacity); support workflows ready and tested; instrumentation in place; and operations support ready.

CPT-502/C: "Computing systems re-visited based on CSA 2006 lessons-learned" (Oct 2006)

The goal of this milestone is to test the operation of the developed and refined computing baseline for 2007. This includes acting on the experience from the Computing, Software and Analysis Challenge (CSA 2006) and the release of pre-final version of components.

CPT-504/C: "Integration of Computing Systems at Tier-0, 1 and 2 centres" (Dec 2006)

The goal of this milestone is to get the computing system for start of data taking deployed and running at the initial set of CMS Tier-0/1/2 centres. This includes:

- integration of Tier-1 and Tier-2 systems
- Data Management tools for data taking in place and tested
- database deployment scheme and workflows understood

The deliverables include: functional and integrated system of Tier-1 and Tier-2; data management tools usable for physics groups; workload management systems ready to

deal with at-scale number of jobs at Tier-2 centers; and database deployment scheme and workflows understood.

CPT-601/C: "Tier-0 centre and CERN Analysis Facility ready for pilot run" (Feb 2007)

The goal of this milestone is to get the CMS-Tier-0 and CMS-CAF ready for the pilot run. It includes the HLT - CMS-Tier-0 integration test and the final version of components integrated and (largely) frozen for startup.

CPT-602/C: "Tier-1 and Tier-2 centres ready for pilot run" (Apr 2007)

The goals of this milestone are the release and test of the distributed environment for the pilot run, including non-CERN Computing Centres.

5.4 Computing Project Resources

The resources associated to the project include the following:

- **Computing capacity requirements** These are given in the next section.
- **General Computing service requirements** These are requirements for services on top of the raw capacities (e.g., database services, batch systems, and user support). These follow from the technical baseline as described in the previous chapter and the regional centre functional description as given above.

  The resources to provide these services at regional centres and at CERN are covered in the draft LCG Computing Grid Collaboration Memorandum of Understanding.
- **CMS-specific Computing service requirements** We have some manpower requirements for CMS-specific services on top of the basic services (e.g., making sure CMS databases have meaningful data in them, running CMS production, skims, and tracking down failures due to CMS application errors) This is *not* part of LCG MoU and is covered by the CMS M&O MoU.

The Computing Task requires Computing Professionals. These people are skilled in areas such as OO analysis and design, C++ and other computing languages, databases and data management systems, computing systems, software process, quality control, and so on. In general, such people have formal computing education and experience although some may be physicists who have changed career path by learning the requisite skills.

Fundamentally, the CMS Computing Environment will be "build to cost", at costs given by the profile of available manpower. The significant implication of this is that the main contingency available to the project is scope contingency. A reduction of scope due to a lack of manpower can have significant implications for the physics program of CMS.

5.4.1 Input Parameters of the Computing Model

We include in Table 5.2 a list of input parameters that have been used in our model calculations for 2008.

5.4.2 Profile of computing resources

Following the same form of calculation as performed in the Computing Model paper we have estimated the time profile for required computing resources. We use the year 2008 as the reference year, and apply some corrections to 2007 and later years based on that reference year. We have made a number of assumptions and simplifications:

2007. Run is assumed to be approximately 50 days of running. The assumption is that LHC duty cycle will be such that it will be possible to keep up with data taking with order half the computing capacity but the storage requirements would be about 1/3 of that for the next run. This data will be very important for detector understanding, but is not likely to be of long-term importance once the data from the initial full run of LHC in 2008 starts to become available.

2008. Numbers are essentially as given in Computing Model paper. No explicit account has been taken of the additional storage/processing of the 2007 sample (except that the tape recorded in 2007 is maintained).

2009. The Event Size assumed to reduce to 1 MB, data rate stays at 225 MB/s. Tier-1 analysis needs double (due to sample size), reprocessing passes reduced to 1 pass, but over twice as much data, thus constant. Disk increases by 50% (only assume one year's raw data disk resident). WAN scaled up by 50%. Tier-2 analysis needs double, simulation assumed constant.

2010. High Luminosity, raw processing time will be up by a factor of 5. Event data volume will be constant. The same event rate as 2009. Tier-0 processing not in real time but using full 200 days. Tier-0 buffer increased to account for the increased risk of falling behind. Tier-1 reprocessing CPU increase by a factor 2.5 (with respect to 2008). (Now only one full reprocessing worldwide possible.) Tier-2 analysis needs increase another 50%, simulation time at high lumi doubles.

- The CMS CERN Analysis Facility (CMS-CAF), which has a special role due to its temporal and geographic closeness to the running experiment, is calculated

Name	Description	Value	Units
L2Rate	pp Rate to Tape	150	Hz
HIRate	Weighted mean HI event Rate	50	Hz
LHCYear	Days of pp Running/year	10000000	sec
HIYear	Seconds of HI Running/year	1.E+06	sec
NRawEvts	Number of pp Raw Events/year	1.5E+09	(derived)
NHIEvts	Number of HI Events/year	5.0E+07	(derived)
RawSize	Raw Data Event Size	1.5	MB
SimSize	Simulated Event Size	2	MB
RecSimSize	Reconstructed Sim Event Size	0.4	MB
RECOSize	Reco Size	0.25	MB
AODSize	AOD Size	0.05	MB
TAGSize	Tag and DPD Size	0.01	MB
HIRawSize	Weighted Mean Heavy Ion Raw Event Size	7	MB
HIRecoSize	Weighted Mean Heavy Ion Reco Size	1	MB
HIAODSize	Weighted Mean Heavy Ion AOD Size	0.2	MB
NPhys	Number of Active Physicists	1000	
NTier1	Number of Tier1 Centers	7	
NTier2	Number of Tier2 Centers	25	
NSimEvt	Number of Simulated Events	1.5.E+09	Evts/Year
FracSimT1	Fraction of NSimEvts done at T1	0%	
NSimPrivate	Number of Private Sim at T2s	8.E+08	Evts/Year
RecCPU	Reconstruction time (Raw)	25	kSI2k.s/ev
SimCPU	Simulation time	45	kSI2k.s/ev
SelCPU	Selection time	0.25	kSI2k.s/ev
AnaCPU	Analysis time	0.25	kSI2k.s/ev
HICPU	Heavy Ion reconstruction time	200	kSI2k.s/ev
NStreamsOFFL	Number of Streams from the off-line	50	
T1RAWCopies	RAW Copies at T1 centers	1	
T0RAWCopy	RAW Copies at CERN	1	
T1RECOCopies	RECO/ESD Copies at T1 Centers	1	
T1AODCopies	AOD Copies	7	
NRECOyear	Reprocessings per year	3	
CalCPU	CPU per Calibration Evt	10	kSI2k.s/ev
CalFrac	Calibration data fraction	10%	
EffSchedCPU	Efficiency factor for Scheduled CPU	85%	
EffAnalCPU	Efficiency factor for Chaotic CPU	75%	
EffDisk	Disk Utilization Efficiency	70%	
EffActiveTape	Active Tape Efficiency	100%	
UserDisk	Group and User Analysis Space	1.0	TB

Table 5.2: Input Parameters for the computing resource calculations for year 2008.

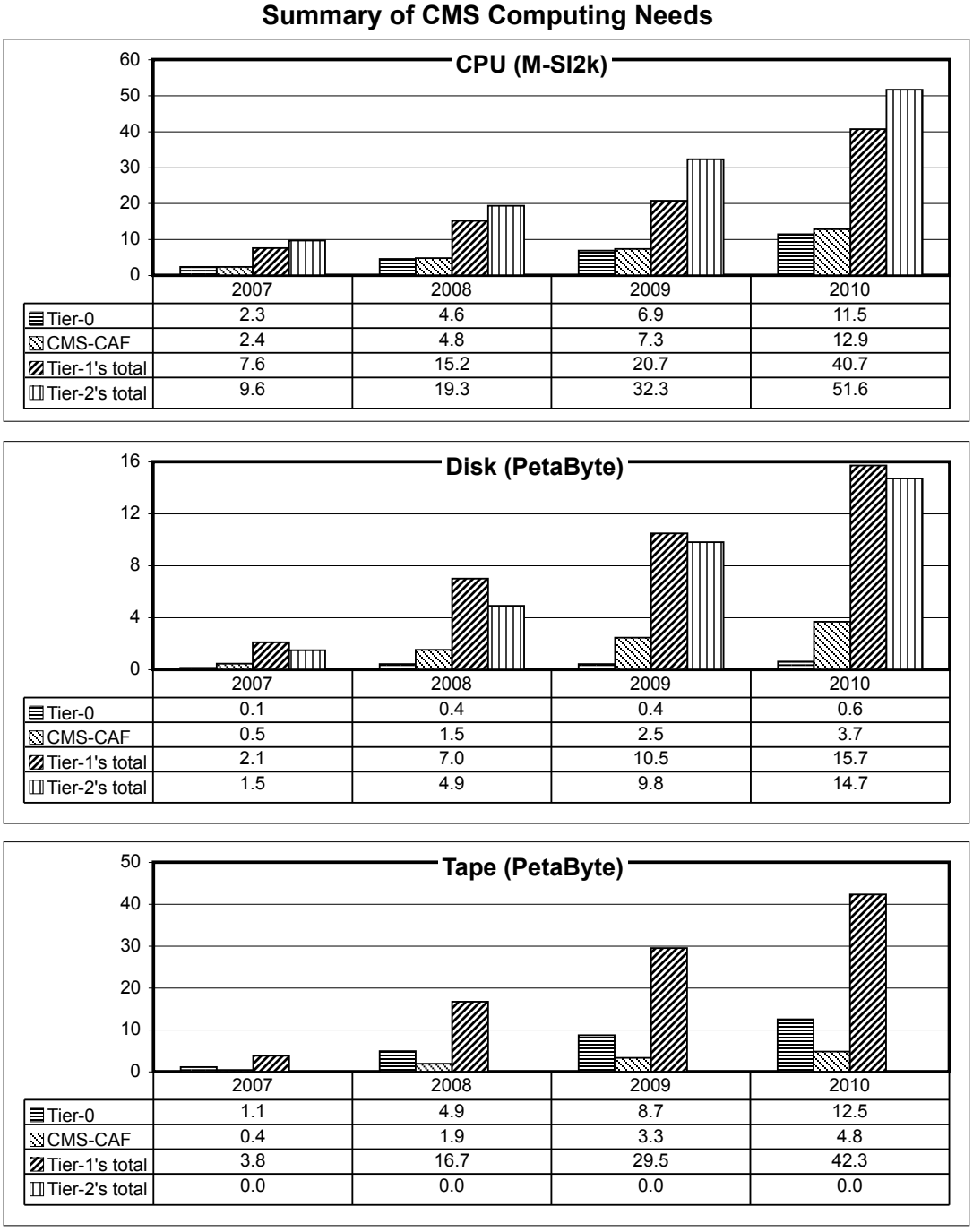
as a standard Tier-1, having taking into account that the raw data is already on tape at CERN Tier-0, plus an additional 2.5 standard Tier-2s to allow for the analysis-like activities at the CMS-CAF.

- Resources of Tier-1s and Tier-2s outside CERN are integrated. We anticipate that CMS will make use of 7-10 physical Tier-1 centres, and 20-25 physical Tier-2 centres.
- Note that WAN calculations in this table do not include factors to account for effective bandwidth usage, whereas those quoted in the Computing Model did include a factor of two for this.

Table 5.3 and Figure 5.3 show the current understanding of the time profile of CMS Computing Requirements.

		Running Year				
		2007	2008	2009	2010	
Conditions		Pilot	2E33+HI	2E33+HI	E34+HI	
Tier-0	CPU	2.3	4.6	6.9	11.5	MSi2k
	Disk	0.1	0.4	0.4	0.6	PB
	Tape	1.1	4.9	9	12	PB
	WAN	3	5	8	12	Gb/s
A Tier-1	CPU	1.3	2.5	3.5	6.8	MSi2k
	Disk	0.3	1.2	1.7	2.6	PB
	Tape	0.6	2.8	4.9	7.0	PB
	WAN	3.6	7.2	10.7	16.1	Gb/s
Sum Tier-1	CPU	7.6	15.2	20.7	40.7	MSi2k
	Disk	2.1	7.0	10.5	15.7	PB
	Tape	3.8	16.7	29.5	42.3	PB
A Tier-2	CPU	0.4	0.9	1.4	2.3	MSi2k
	Disk	0.1	0.2	0.4	0.7	PB
	WAN	0.3	0.6	0.8	1.3	Gb/s
Sum Tier-2	CPU	9.6	19.3	32.3	51.6	MSi2k
	Disk	1.5	4.9	9.8	14.7	PB
CMS CERN	CPU	2.4	4.8	7.3	12.9	MSi2k
Analysis Facility	Disk	0.5	1.5	2.5	3.7	PB
(CMS-CAF)	Tape	0.4	1.9	3.3	4.8	PB
	WAN	0.3	5.7	8.5	12.7	Gb/s
Total	CPU	21.9	43.8	67.2	116.6	MSi2k
	Disk	4.1	13.8	23.2	34.7	PB
	Tape	5.4	23.4	41.5	59.5	PB

Table 5.3: Time Profile of CMS Computing Requirements



	2007	2008	2009	2010
Tier-0	2.3	4.6	6.9	11.5
CMS-CAF	2.4	4.8	7.3	12.9
Tier-1's total	7.6	15.2	20.7	40.7
Tier-2's total	9.6	19.3	32.3	51.6

	2007	2008	2009	2010
Tier-0	0.1	0.4	0.4	0.6
CMS-CAF	0.5	1.5	2.5	3.7
Tier-1's total	2.1	7.0	10.5	15.7
Tier-2's total	1.5	4.9	9.8	14.7

	2007	2008	2009	2010
Tier-0	1.1	4.9	8.7	12.5
CMS-CAF	0.4	1.9	3.3	4.8
Tier-1's total	3.8	16.7	29.5	42.3
Tier-2's total	0.0	0.0	0.0	0.0

Figure 5.3: Time Profile of CMS Computing Requirements

Appendix A

Requirements and Specifications from the Computing Model Paper

In this appendix we recall the Requirements and Specifications identified in the Computing Model Paper [2]

A.1 Requirements

R-1 The online HLT system must create "RAW" data events containing: the detector data, the L1 trigger result, the result of the HLT selections ("HLT trigger bits"), and some of the higher-level objects created during HLT processing.

R-2 The RAW event size *at startup* is estimated to be $S_{RAW} \simeq 1.5$ MB, assuming a luminosity of $\mathcal{L} = 2 \times 10^{33}\text{cm}^{-2}\text{s}^{-1}$.

R-3 The RAW event size *in the third year of running* is estimated to be $S_{RAW} \simeq 1.0$ MB, assuming a luminosity of $\mathcal{L} = 10^{34}\text{cm}^{-2}\text{s}^{-1}$.

R-4 The RAW event rate from the online system is 150 Hz or 1.5×10^9 events per year.

R-5 Event reconstruction shall generally be performed by a central production team, rather than individual users, in order to make effective use of resources and to provide samples with known provenance and in accordance with CMS priorities.

R-6 CMS production must make us of data provenance tools to record the detailed processing of production datasets and these tools must be useable (and used) by all members of the collaboration to allow them also this detailed provenance tracking

R-7 The reconstructed event format (RECO) is about 250 KByte/event; it includes quantities required for all the typical analysis usage patterns such as: pattern recognition in the tracker, track re-fitting, calorimeter re-clustering, and jet energy calibration.

R-8 The AOD data format at low luminosity shall be approximately 50kB/event and contain physics objects: tracks with corresponding RecHit's, calorimetric clusters with corresponding RecHit's, vertices (compact), and jets.

R-9 The data rate (MB/s) for Heavy Ion running will be approximately the same as that of pp running however event sizes will be substantially higher, around 5-10MB/event.

R-10 Heavy Ion events will be reconstructed during the (approximately 4 month) period between LHC operations periods; it is not necessary to keep up with data taking as for pp running.

R-11 Heavy Ion reconstruction is costly, 10-50 times that of pp reconstruction. The base Heavy Ion program (as in the CMS Technical Proposal) can be achieved with the lower number, more physics can be reached with the higher.

R-12 The online system shall temporarily store "RAW" events selected by the HLT, prior to their secure transfer to the offline Tier-0 centre.

R-13 The online system will classify RAW events into $\mathcal{O}(50)$ Primary Datasets based solely on the trigger path (L1+HLT); for consistency, the online HLT software will run to completion for every selected event.

R-14 For performance reasons, we may choose to group sets of the $\mathcal{O}(50)$ Primary Datasets into $\mathcal{O}(10)$ "Online Streams" with roughly similar rates.

R-15 The Primary Dataset classification shall be immutable and only rely on the L1+HLT criteria which are available during the online selection/rejection step.

R-16 Duplication of events between Primary Datasets shall be supported (within reason - up to about approximately 10%).

R-17 The online system will write one or possibly several "Express-Line" stream(s), at a rate of a few % of the total event rate, containing (by definition) any events which require very high priority for the subsequent processing.

R-18 The offline system must be able to keep up with a data rate from the online of about 225 MB/s. The integrated data volume that must be handled assumes 10^7 seconds of running.

R-19 No TriDAS dead-time can be tolerated due to the system transferring events from the online systems to the Tier-0 centre; the online-offline link must run at the same rate as the HLT acceptance rate.

R-20 The primary data archive (at the Tier-0) must be made within a delay of less than a day so as to allow online buffers to be cleared as rapidly as possible

R-21 CMS requires an offline first-pass full reconstruction of express line and all online streams in quasi-realtime, which produces new reconstructed objects called RECO data.

R-22 A crucial data access pattern, particularly at startup will require efficient access to both the RAW and RECO parts of an event

R-23 The reconstruction program should be fast enough to allow for frequent reprocessing of the data.

R-24 Fully simulated Monte Carlo samples of approximately the same total size as the raw data sample (1.5×10^9 events per year) must be generated, fully simulated, reconstructed and passed through HLT selection code. The simulated pp event size is approximately 2 MByte/event.

R-25 Fully simulated Monte Carlo samples for Heavy Ion physics will be required, although the data volume is expected to be modest compared to the pp samples.

R-26 CMS needs to support significant amounts of expert analysis using RAW and RECO data to ensure that the detector and trigger behaviour can be correctly understood (including calibrations, alignments, backgrounds, etc.).

R-27 Physicists will need to perform frequent skims of the Primary Datasets to create sub-samples of selected events.

R-28 CMS needs to support significant physics analysis using RECO and AOD data to ensure the widest range of physics possibilities are explored.

R-29 The AOD data shall be the primary event format made widely available for physics analysis in CMS.

R-30 Access to information stored in AOD format shall occur through the same interfaces as are used to access the corresponding RECO objects.

R-31 An "Event directory" system will be implemented for CMS.

R-32 Smaller and more specialised TAG/tuple data formats can be developed as required.

R-33 Multiple GRID implementations are assumed to be a fact of life. They must be supported in a way that renders the details largely invisible to CMS physicists.

R-34 The GRID implementations should support the movement of jobs and their execution at sites hosting the data, as well as the (less usual) movement of data to a job. Mechanisms should exist for appropriate control of the choices according to CMS policies and resources.

A.2 Specifications

S-1 The link from the online to the Tier-0 centre should be sized to keep up with the event flow from the Online farm, with an additional safety margin to permit the clearing of any backlogs caused by downstream throughput problems in the Tier-0.

S-2 The processing capacity of the Tier-0 centre should be sufficient to keep up reconstructing the real-time event flow from the CMS online system.

S-3 The RAW and RECO data components (i.e. the FEVT) of a given set of events are, by default, distributed together. The technical ability to ship them seperately should the need arise shall be maintained.

S-4 The first pass reconstruction step will also produce the AOD data, a copy of which is sent to every single Tier-1 Centre.

S-5 Two copies of the CMS RAW Data shall be kept on long term secure storage media (tape): one copy at the Tier 0 and a second copy at the ensemble of Tier-1 centres.

S-6 The Tier-0 shall store all CMS RAW data on secure storage media (tape) and maintain it long-term.

S-7 The Tier-0 centre shall store a secure copy of all data which it produces as part of its official CMS production passes, including first pass reconstruction (RECO) output, subsequent re-processing steps, and any AOD's produced.

S-8 Tape Storage at the Tier-0 and Tier-1 centres shall be used as a trusted archive and an active tertiary store

S-9 The Tier-0 storage and buffer facility shall be optimised for organised and scheduled access during experimental running periods

S-10 The Tier-0 will not support logins from general CMS users, only those carrying out specific production related activities.

S-11 The Tier-0 shall support at least one complete re-reconstruction pass of all RAW data, using calibrations and software which are improved compared to the original first-pass processing.

S-12 The re-reconstruction step will also produce the AOD data, a copy of which is sent to every single Tier-1 Centre.

S-13 About half of the Tier-0 capacity could be used to perform regional reconstruction of Heavy Ion events during LHC downtimes. This time does however eat into that available for re-reconstruction

S-14 CPU resources at some Tier-2 centres could be used to carry out the Heavy ion initial reconstruction, or to extend that reconstruction to allow more physics coverage

S-15 The Tier-0 shall coordinate the transfer of each Primary Dataset in FEVT format, and all AOD data produced, to a "custodial" Tier 1 centre prior to its deletion from the Tier-0 output buffer.

S-16 The Tier-0 centre shall support a range of collaboration services such as: resource allocation and accounting, support for CMS policies; high- and low-level monitoring; data catalogs; conditions and calibration databases; software installation and environment support; virtual organisations and other such services.

S-17 The ensemble of non-CERN Tier-1 centres shall store the second "custodial" copy of the FEVT (= RAW + RECO) data coming from the Tier-0, on secure storage media

(tape) and maintain it long-term.

S-18 The Tier-1's must have sufficient processing resources to re-reconstruct the RAW data entrusted to that centre twice per year, in addition to the single full reprocessing at the Tier-0 during the LHC shutdowns.

S-19 The Tier-1's must have sufficient processing resources to re-process (reconstruct) twice per year the MC samples which they host .

S-20 Tier-1 centres must store a secure copy of all data they produce as part of official CMS production passes, including RECO and AOD formats.

S-21 Tier-1 centres shall support limited interactive and batch analysis of data which they host.

S-22 Tier-1 centres shall support massive selection and skim passes through the data that they host and distribute the product datasets to the requesting Tier-2 centres

S-23 Tier-1 centre selection facilities will require high performance (order 800MB/s) data-serving capacity from their local data samples to their selection farms

S-24 Tier-1 centres must offer sufficiently granular job submission queues to enable CMS to partition priorities arbitrarily between (perhaps different) analysis groups and individuals

S-25 Each of the $(N_{T1}-1)$ Tier-1 centres must size its network to: accept its $\sim 1/(N_{T1}-1)$ share of total RAW and RECO data produced at the Tier-0 during running periods; accept MC production data from $\sim N_{T2}/N_{T1}$ of the N_{T2} Tier-2 centres; and export requested datasets to $\sim N_{T2}/N_{T1}$ Tier-2 regional centres.

S-26 CMS requires Tier-1 functionality at CERN

S-27 Some portion of the Raw + Reconstructed data will be served from the Tier-1 centre at CERN, but the full second copy of the data will be spread across the regional Tier-1 centres.

S-28 Tier-1 centres shall support a range of collaboration services such as: resource allocation and accounting, support for CMS policies; high- and low-level monitoring; data catalogs; conditions and calibration databases; software installation and environment support; virtual organisations and other such services.

S-29 Tier-2 centres shall dedicate a significant fraction of their processing capacity to their associated analysis communities.

S-30 Tier-2 centres should have WAN connectivity in the range of 1Gb/s or more to satisfy CMS analysis requirements

S-31 Tier-2 centres will require relatively sophisticated disk cache management systems, or explicit and enforceable local policy, to ensure sample latency on disk is adequate and to avoid disk/WAN thrashing

S-32 Tier-2 centres should provide processing capacity for the production of standard CMS Monte Carlo samples ($\sim 10^9$ events/year summed over all centres), including full detector simulation and the first pass reconstruction.

S-33 Some Tier-2 centres will provide processing power to allow the Heavy Ion reconstruction to be completed, or extended compared to that available at the Tier-0

S-34 CMS requires Tier-2 functionality at CERN

S-35 Tier-2 centres are responsible for guaranteeing the transfer of the MC samples they produce to a Tier-1 which takes over custodial responsibility for the data.

S-36 Tier-2 computing centres have no custodial responsibility for any data.

Appendix B

Further Reading

Physics Software:

- OSCAR: An Object-Oriented Simulation Program for CMS [32]
- FAMOS: a FAst MOnte Carlo Simulation for CMS [33]
- Mantis: a Framework and Toolkit for Geant4-Based Simulation in CMS [34]
- CMKIN v3 User's Guide [35]
- ORCA: reconstruction program [36, 37, 38, 39]
- Magnetic field software implementation in CMS [40]
- High Level Trigger software for the CMS experiment [41]
- Monitoring CMS Tracker construction and data quality using a grid/web service based on a visualization too [42]
- Expected Data Rates from the Silicon Strip Tracker [10]

DC04 Data Challenge (computing aspects):

- Distributed Computing Grid Experiences in CMS DC04 [43]
- Role of Tier-0, Tier-1 and Tier-2 Regional Centers during CMS DC04 [44]
- Tier-1 and Tier-2 Real-time Analysis experience in CMS DC04 [45]
- Production Management Software for the CMS Data Challenge [46]
- Planning for the 5% Data Challenge, DC04 [47]
- CMS Distributed Data Analysis Challenges [48]
- Distributed File system Evaluation and Deployment at the US-CMS Tier-1 Center [49]
- Software agents in data and workload management [16]

DC04 Data Challenge (analysis experiences):

- Using the reconstruction software, ORCA, in the CMS data challenge 2004 [50]
- Use of Grid Tools to Support CMS Distributed Analysis [51]
- Distributed Computed Grid Experiences in CMS DC04 [43]
- Grid Enabled Analysis for CMS: prototype, status and results [52]
- GROSS: an end user tool for carrying out batch analysis of CMS data on the LCG-2 Grid. [53]
- Clarens Web services [54]

Production systems:

- RefDB (the Reference Database for CMS Monte Carlo Production) [55, 31]
- McRunjob (a High Energy Physics Workflow Planner for Grid Production Processing) [56]
- BOSS (an Object Based System for Batch Job Submission and Monitoring) [19]
- Virtual Data in CMS Production [57]
- Combined Analysis of GRIDICE and BOSS Information Recorded During CMS-LCG0 Production [58]
- Running CMS Software on GRID Testbeds [59]
- Resource Monitoring Tool for CMS production [60]
- The Spring 2002 DAQ TDR Production [61]
- CMS Test of the European DataGrid Testbed [62]
- Use of Condor and GLOW for CMS Simulation Production [63]
- Study and Prototype Implementation of a Distributed System [64]

Core Applications Software:

- Report of the CMS Data Management RTAG [65]
- Status and Perspectives of Detector Databases in the CMS Experiment at the LHC [66]
- Modeling a Hierarchical Data Registry with Relational Databases in a Distributed Environment [67]
- Detector Geometry Database [68]
- Migration of the XML Detector Description Data and Schema to a Relational Database [69]

- De-serializing Object Data while Schemas Evolve [70]
- Evaluation of Oracle9i C++ Call Interface [71]
- 3D Graphics Under Linux [72]
- IGUANA Plan For 2002 [73]
- Evaluation Of Oracle9i To Manage CMS Event Store: Oracle Architecture To Store Petabyte Of Data (PART ONE) [74]
- Composite Framework for CMS User Applications [75]
- Mantis: the Geant4-based simulation specialization of the CMS COBRA framework [76]
- CMS Detector Description: New Developments [77]
- A database perspective on CMS data [78]
- ROOT - An Object Oriented Data Analysis Framework [79]

Software Environment:

- Use Cases and Requirements for Software Installation in Grid and End-User Desktop Environments [80]
- OVAL: The CMS Testing Robot [81]
- Installation/Usage Notes For Oprofile [82]
- CMS Software Quality [83]
- Evaluation Of The CMT And SCRAM Software Configuration, Build And Release Management Tools [84]
- CMS Software Installation [23]
- Parallel compilation of CMS software [85]
- PRS Software Quality Policy [86]
- Software Metrics Report Of CMS Reconstruction Software [87]

General Organisation and Planning:

- CMS Computing and Software Tasks and Manpower for 2003-2007 [88]
- Computing And Core Software (CCS) Schedule And Milestones: Version 33 [89]
- Planning for CTDR [90]
- Proposed Scope And Organization Of CMS-CPT. Computing And Core Software, Physics Reconstruction and Selection, TriDAS (Online Computing) [91]
- CMS Grid Implementation Plan - 2002 [92]

- Plans for the Integration of Grid Tools in the CMS Computing Environment [93]
- Scope and Organization of CMS-CPT [94]

Computing at the Tevatron

- Job and Information Management Deployment for the CDF Experiment [95]
- Monitoring the CDF distributed computing farms [96]
- Testing the CDF Distributed Computing Framework [97]
- Tools for GRID deployment of CDF offline and SAM data handling systems for Summer 2004 computing [98]
- Globally Distributed User Analysis Computing at CDF [99]
- Deployment of SAM for the CDF Experiment [100]
- The Condor based CDF CAF [101]
- Performance of an operating High Energy Physics Data grid, D0SAR-grid [102]
- D0 data processing within EDG/LCG [103]
- Experience using grid tools for CDF physics [104]

Appendix C

Computing Project Participants

Yerevan Physics Institute, Yerevan, ARMENIA
S. Chatrchyan and A.M. Sirunyan

Université Catholique de Louvain, Louvain-la-Neuve, BELGIUM
G. Bruno, C. Delaere, P. Demin, T. Keutgen, V. Lemaitre, A. Ninane and O. van der Aa

Université de Mons-Hainaut, Mons, BELGIUM
A. Romeyer

Vrije Universiteit Brussel, Brussel, BELGIUM
S. De Weirdt, S. Lowette and S. Rugovac

Université Libre de Bruxelles, Bruxelles, BELGIUM
O. Bouhali, P. Vanlaer and S. Viji

University of Sofia, Sofia, BULGARIA
N. Darmenov, L. Litov, Z. Toteva and V. Verguilov

National Institute of Chemical Physics and Biophysics, Tallinn, ESTONIA
A. Hektor and M. Kadastik

Helsinki Institute of Physics, Helsinki, FINLAND
V. Karimäki, J. Klem and T. Linden

Institut de Physique Nucléaire de Lyon, IN2P3-CNRS, Université Lyon I, Villeurbanne, FRANCE
D. Bouvet, J. Devémy, P. Gaillardon, F. Hernandez, T. Kachelhoffer, N. Lajili, J.-Y. Nief, S. Poulat and L. Schwartz

Laboratoire Leprince-Ringuet, Ecole Polytechnique, IN2P3-CNRS, Palaiseau, FRANCE
C. Charlot, A-M. Gaillac, P. Miné and I. Semeniouk

Institut für Experimentelle Kernphysik, Karlsruhe, GERMANY
C. Jung, T. Muller, G. Quast, K. Rabbertz, J. Rehn[**a], A. Schmidt, A. Vest, C. Weiser and J. Weng[**b]

RWTH, III. Physikalisches Institut B, Aachen, GERMANY
M. Duda, M. Erdmann, M. Kirsch, T. Kress and A. Nowack

Università di Bari, Politecnico di Bari e Sezione dell' INFN, Bari, ITALY
N. De Filippis, D. Diacono, G. Donvito, R. Gervasoni, G. Maggi, M. Maggi, M. Mennea. A. Pierro, L. Silvestris and G. Zito

Università di Bologna e Sezione dell' INFN, Bologna, ITALY
D. Bonacorsi, P. Capiluppi, A. Fanfani, C. Grandi and A. Sciabà

Università di Catania e Sezione dell' INFN, Catania, ITALY
S. Costa and A. Tricomi

Università di Firenze e Sezione dell' INFN, Firenze, ITALY
G. Ciraolo, V. Ciulli and N. Magini

Laboratori Nazionali di Legnaro dell' INFN, Legnaro, ITALY (associated institute)
S. Badoer, L. Berti, M. Biasotto and G. Maron

Istituto Nazionale di Fisica Nucleare e Universita Degli Studi Milano-Bicocca, Milano, ITALY
M. Bonesini, F. Ferri, P. Govoni and M. Paganoni

Università di Padova e Sezione dell' INFN, Padova, ITALY
S. Fantinel, F. Fanzago, S. Lacaprara, M. Mazzucato and N. Smirnov

Università di Perugia e Sezione dell' INFN, Perugia, ITALY
F. Ambroglini, L. Faina, L. Fanò, M. Mariotti, L. Servoli and D. Spiga

Università di Pisa, Scuola Normale Superiore e Sezione dell' INFN, Pisa, ITALY
G. Bagliesi, T. Boccali, F. Donno, L. Foà, S. Gennai and R. Tenchini

Università di Roma "La Sapienza" e Sezione dell' INFN, Roma, ITALY
L.M. Barone, P. Meridiani and S. Rahatlou

Università di Torino e Sezione dell' INFN, Torino, ITALY
N. Amapane[**a], R. Bellan and G. Cerminara

INFN - Sezione di Trieste, Trieste, ITALY
S. Belforte

National Centre for Physics, Quaid-I-Azam University, Islamabad, PAKISTAN
U. Ahmad, S. Asghar and M. Hafeez

Pakistan Atomic Energy Commission, Islamabad, PAKISTAN (related institute)
N. Batool, A.A. Huqqani, A. Mahmood and A. Osman

Institute for Nuclear Research, Moscow, RUSSIA
M. Kirsanov and A. Toropin

Institute for Theoretical and Experimental Physics, Moscow, RUSSIA
V. Gavrilov, N. Ilyina, E. Lyublev, V. Kolossov, A. Krokhotin and A. Oulyanov

Joint Institute for Nuclear Research, Dubna, RUSSIA
I. Belotelov, I. Filozova, I. Golutvin, V. Konoplyanikov, V. Korenkov, A. Lanyov, V. Mitsyn, P. Moissenz, E. Nikonov, D. Oleynik, V. Palichik, A. Petrosyan, E. Rogalev, R. Semenov, S. Shulga, S. Shmatov and E. Tikhonenko

P.N. Lebedev Physical Institute, Moscow, RUSSIA
I. Dremin, N. Konovalova and N. Lvova

Moscow State University, Moscow, RUSSIA
A. Berejnoi, A. Demichev, L. Dudko, V. Edneral, A. Gribushin, V. Ilyin, O. Kodolova, N. Kruglov, A. Kryukov, I. Lokhtin, A. Snigirev, L. Shamardin, A. Sherstnev, G. Spiez, S. Petrushanko, K. Teplov, I. Vardanyan and S. Zotkin

Petersburg Nuclear Physics Institute, Gatchina (St Petersburg), RUSSIA
A. Grebenyuk, J. Grebenyuk, V. Kim, A. Kiryanov, D. Kozlenko and A. Oreshkin

State Research Center of Russian Federation - Institute for High Energy Physics, Protvino, RUSSIA
S. Bityukov, K. Datsko, A. Filine, D. Konstantinov, Yu. Lazin, V. Petoukhov, V. Petrov, R. Ryutin, M. Sapunov, E. Slabospitskaya, S. Slabospitsky, A. Sobol, N. Tyurin and V. Urazmetov

Centro de Investigaciones Energeticas Medioambientales y Tecnologicas, Madrid, SPAIN
M. Cardenas, N. Colino, P. Garcia-Abia, J.M. Hernandez, G. Merino, E. Perez and F.J. Rodriguez-Calonge

Instituto de Física de Cantabria (IFCA), CSIC-Universidad de Cantabria, Santander, SPAIN
D. Cano, I. Gonzalez-Caballero, J. Marco, R. Marco, C. Martínez-Rivero, F. Matorras and D. Rodríguez

Universidad de Oviedo, Oviedo, SPAIN
J. Cuevas, J.M. Lopez and H. Naves

CERN, European Organization for Nuclear Research, Geneva, SWITZERLAND
S. Argiro, S. Ashby, M. Corvo, N. Darmenov[**c], A. De Roeck, J. Herrala, V. Innocente, W. Jank, P. Janot, V. Karimaki[**d], V. Lefebure[**e], M. Liendl[**e], E. Meschi, F. Moortgat, C. Munro[**f], M. Pimia, J-P. Porte[**e], J. Rehn, D. Samyn, N. Sinanis, P. Sphicas, M. Spiropulu, H. Stockinger[**j], Z. Toteva[**c], R. Voicu[**g] and I. Willers[**i]

CERN, European Organization for Nuclear Research, Geneva, SWITZERLAND and INFN, Istituto Nazionale di Fisica Nucleare, ITALY
S. Argiro and M. Corvo

National Scientific Center, Kharkov Institute of Physics and Technology, Kharkov, UKRAINE
D. Soroka, S. Zub and L. Levchuk

Brunel University, Uxbridge, UNITED KINGDOM
P.R. Hobson, P. Kyberd and J.J. Nebrensky

Imperial College, University of London, London, UNITED KINGDOM
D. Colling, O. Maroney, B. MacEvoy, H. Tallini, S. Wakefield and Y. Zhang

Rutherford Appleton Laboratory, Didcot, UNITED KINGDOM
T. Folkes, D. Ross, A. Sansum, C.H. Shepherd-Themistocleous, B. Strong, S. Traylen and N. White

University of Bristol, Bristol, UNITED KINGDOM
T. Barrass, S. Metson and D. Newbold[**h]

Centre for Complex Cooperative Systems, University of the West of England, Bristol, UNITED KINGDOM (associated institute)
A. Anjum, N. Baker, F. Estrella, W. Hassan, T. Hauer, D. Manset, R.H. McClatchey, D. Rogulin and A.E. Solomonides

Boston University, Boston, Massachusetts, USA
E. Hazen, A.H. Heering, D. Lazic, J. Rohlf, F. Varela and S.X. Wu

California Institute of Technology, Pasadena, California, USA
P. Angelino[**e], T. Azim, J. Bunn, J. Choi, P. Galvez, S. Iqbal, I. Legrand, J. Lindheim, V. Litvine, P. Messina[**e], D. Nae, H.B. Newman, S. Ravot, S. Singh, C. Steenberg, S. Singh, X. Su, M. Thomas, F. Van Lingen, R. Wilkinson and X. Yang

Fermi National Accelerator Laboratory, Batavia, Illinois, USA
A. Afaq, J. Amundson, W. Baisley, J. Bakken, L.A.T. Bauerdick, W. Brown, T. Doody, D. Evans, D. Fagan, I. Fisk, L. Giacchetti, G. Graham, G. Guglielmo, Y. Guo, R.M. Harris, J. Kaiser, M. Leininger, L. Lueking, T. Levshina, J. Marraffino, T. Messer, V. O'Dell, M. Paterno, T. Perelmutov, D. Petravick, R. Pordes, N. Ratnikova, V. Sekhri, E. Sexton-Kennedy, D. Skow, G. Stiehr, M. Stavrianakou, W. Tanenbaum, H. Wenzel, E. Wicklund, V. White, Y. Wu and A. Yagil

Northeastern University, Boston, Massachusetts, USA
G. Alverson, G. Eulisse, S. Muzaffar, I. Osborne, L. Taylor and L. Tuura

Princeton University, Princeton, New Jersey, USA
P. Elmer, D. Stickland, C. Tully, A. Wildish, S. Wynhoff and Z. Xie

Purdue University, West Lafayette, Indiana, USA
N. Neumeister

University of California, Davis, Davis, California, USA
M. Case and P.T. Cox

University of California, Riverside, Riverside, California,USA
J. Andreeva[**a] and R. Clare

University of California, San Diego, La Jolla, California, USA
J. Branson, S.-C. Hsu, J. Letts, E. Lipeles, T. Martin, M. Norman, A. Rana and F. Würthwein

University of Florida, Gainesville, Florida, USA
D. Acosta, P. Avery, D. Bourilkov, R. Cavanaugh, Y. Fu, B. Kim, C. Prescott and J. Rodriguez

University of Wisconsin, Madison, Wisconsin, USA
D. Bradley and S. Dasu

**a: Now at CERN.

**b: Also at CERN.

**c: Also at Bulgaria.

**d: Also at Helsinki.

**e: No longer affiliated with CMS.

**f: Also at Brunel University.

**g: Also with the EU.

**h: Also at RAL.

**i: Also at UWE.

**j: Now at Vienna.

Appendix D

Glossary

AFS	Andrew File System
ANSI	American National Standards Institute
AOD	Analysis Object Data - a compact event format for physics analysis
API	Application Programming Interface
ATM	Asynchronous Transfer Mode
BOSS	Object Based System for Batch Job Submission and Monitoring
CAD	Computer-Aided Design
CAF	CERN Analysis Facility
CASE	Computer-Aided Software Engineering
CASTOR	CERN Advanced Storage Manager
CCC	Computing Coordination Committee
CD	Compact Disk
CDF	Collider Detector Facility experiment at the FNAL Tevatron
CDR	Central Data Recording
CE	Computing Element
CIM	Common Information Module
CLHEP	Class Library for HEP
CM	Computing Model
CMKIN	CMS Kinematics Package (legacy Fortran)
CMS	Compact Muon Solenoid
CMSIM	CMS Simulation Package (legacy Fortran)
COBRA	Coherent Object-oriented Base for Reconstruction, Analysis and simulation (Framework)
CODEC	Compression/Decompression
CORBA	Common Object Request Broker Architecture
CPU	Central Processing Unit
CRAB	CMS Remote Analysis Builder
CVS	Concurrent Versions System
DØ	DØ experiment at the FNAL Tevatron
DAG	Directed Acyclic Graph
Dag	OED: noun Austral./NZ a lock of wool matted with dung hanging from the hindquarters of a sheep
DAQ	Data Acquisition
DBMS	Database Management System
DBS	Dataset Book-keeping Service
DCS	Detector Control System
DDL	Data Description Language
DFS	Distributed File System
Digi	Digitisation (of detector hit)

DLS Dataset Location Service
DLT Digital Linear Tape
DM(S) Data Management (System)
DQM Data Quality Manager
DSID Data Set Identifier
DST Data Summary Tape - a compact event format
DVD Digital Versatile Disk

ECAL Electromagnetic Calorimeter
EDG European DataGrid
EDM Event Data Model
EDMS Engineering Database Management System
EGEE Enabling Grids for e-science in Europe (a Grid project)
EFU Event Filter Unit
EPICS Experimental Physics Industrial Control System
ESNET Energy Science Network (in the USA)
EVM Event Manager
Express-Line Online stream for events requiring high priority and low latency offline processing

FAMOS FAst MOnte Carlo Simulation
FDDI Fibre Distributed Data Interface
FE Front-End
FED Front-End Driver
FEVT Event format comprising the union of RAW and RECO data
FNAL Fermi National Accelerator Laboratory, USA
FW Framework

GEANT4 Simulation Framework and Toolkit
GIPS Giga (10^9) Instructions per Second
Gb Gigabit (10^9 bits)
GB Gigabyte (10^9 bytes)
GIF Graphics Interchange Format
GL Graphics Language (low-level 3D rendering software)
GLOW Grid Laboratory of Wisconsin
gPLAZMA grid-aware PLuggable AuthoriZation MAnagement
Grid Infrastructure for Distributed Computing
GridICE Grid Monitoring software from INFN
GUI Graphical User Interface

HCAL Hadronic Calorimeter
HEP High Energy Physics
HEPEVT HEP Event (generated event format)
HEPiX HEP Unix environment
HI Heavy Ion(s)
HLT Higher Level Trigger (Software)
HTML Hypertext Mark-up Language

IGUANA Interactive Graphics for User ANAlysis - used for the CMS Event Display Package
I/O Input/Output
IP Internet Protocol
IPC Interprocess Communication
ISDN Integrated Services Digital Network
IT Information Technology

JDL Job Description Language

kb kilobit (10^3 bits)
kB kilobytes (10^3 bytes)

L1 Level 1 hardware-based trigger
LAN Local Area Network
LCG LHC Computing Grid (a common computing project)
LEP Large Electron Positron Collider
LFC Local File Catalog
LFN Logical File Name
LHC Large Hadron Collider
LHCC LHC (review) Committee

Mb	Megabit (10^6 bits)
MB	Megabyte (10^6 bytes)
MBONE	Multicast Backbone
MC	Monte Carlo simulation program/technique
MCPS	Monte Carlo Production System
MIPS	Mega (10^6) Instructions per Second
MONARC	Models of Networked Analysis at Regional Centres
MS	Microsoft (Corporation)
NQS	Network Queueing System
ODBMS	Object Database Management System
Online-Stream	Grouping of events (Primary Datasets) to simplify online data management
OO	Object Oriented
OQL	Object Query Language
ORB	Object Request Broker
ORCA	CMS Reconstruction Program
OS	Operating System
OSCAR	CMS GEANT4 Simulation Program
OSF	Open Software Foundation
OVAL	CMS Software testing tool
PASTA	LHC Technology Tracking Team
PAW	Physics Analysis Workstation (legacy interactive analysis application
Pb	Petabit (10^{15} bits)
PB	Petabyte (10^{15} bytes)
PFN	Physical File Name
PFS	Physics Reconstruction and Selection
PhEDEx	Physics Experiment Data Export
PNFS	Perfectly Normal File System
POOL	Persistency software from LCG
POSIX	Portable Operating System Interface
Primary-Dataset	Grouping of events according to physics (trigger) criteria
PROOF	Parallel ROOT Facility
QA	Quality Assurance
QC	Quality Control
QOS	Quality of Service
RAID	Redundant Arrays of Independent Disks
RAW	Event format from the online containing full detector and trigger data
RC	Regional Centre / Readout Crate
RECO	Event format for reconstructed objects such as tracks, vertices, jets, etc.
RecHit	Reconstructed hit in a detector element
R-GMA	Relational Grid Monitoring Architecture
RHIC	Relativistic Heavy Ion Collider (at Brookhaven, USA)
RISC	Reduced Instruction Set Computer
RPM	Redhat Package Manager
RRB	Resources Review Board
RTAG	Requirements and Technology Assessment Group
R/W	Read/Write
SAM	Sequential Access via Metadata
SA/SD	Structured Analysis/Structured Design
SCRAM	Software Configuration, Release and Management
SE	Storage Element
SFI	Switch Farm Interface
ShREEK	Shahkar Runtime Execution Environment Kit

Skim Subset of events selected from a larger set
SMP Symmetric Multiprocessor
SNMP Simple Network Management Protocol
SQA Software Quality Assurance
SQL Structured Query Language
SRM Storage Resource Management
STL Standard Template Library
SURL Storage URL

TAG Event index information such as run/event number, trigger bits, etc.
Tb Terabit (10^{12} bits)
TB Terabyte (10^{12} bytes)
TCP Transmission Control Protocol
TDR Technical Design Report
TIPS Tera (10^{12}) Instructions per Second
TMB Thumbnail event format

UI User Interface
UID User ID
URL Uniform Resource Locator

VCAL Very Forward Calorimeter
VO(MS) Virtual Organisation (Membership Service)

WAN Wide Area Network
WLCG Worldwide LHC Computing Grid - being all the resources available to LHC Computing
WM(S) Workload Management (System)
WWW World Wide Web
WN Worker Node
WYSIWYG What You See Is What You Get (type of GUI)

Appendix E

References

Note:
CHEP04 documents on InDiCo are available at
`http://indico.cern.ch/confAuthorIndex.py?confId=0`
CMS Internal Notes are available upon approved request from the CMS Secretariat.
CMS Notes are available at `http://cmsdoc.cern.ch/docall.shtml` unless otherwise noted.

[1] CMS Collaboration, "The Compact Muon Solenoid Computing Technical Proposal," *CERN/LHCC* **1996-045** (1996).

[2] Editors: C. Grandi, D. Stickland, L. Taylor, "The CMS Computing Model," *CMS NOTE* **2004-031** (2004). Also available as CERN LHCC 2004-035/G-083.

[3] M. Aderholz *et al.*, "Models of Networked Analysis at Regional Centres for LHC Experiments (MONARC) - Phase 2 Report," *CERN/LCB* **2000-001** (2000).

[4] S. Bethke *et al.*, "Report of the Steering Group of the LHC Computing Review," *CERN/LHCC* **2001-004** (2001).

[5] I. Bird *et al.*, "Operating the LCG and EGEE Production Grids for HEP," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo (slides; video available).

[6] I. Foster *et al.*, "The Grid2003 Production Grid: Principles and Practice," in *Proceedings of the 13^{th} IEEE Intl. Symposium on High Performance Distributed Computing, 2004*. Honolulu, Hawaii, June 4^{th}-6^{th} 2004, 2004. Available at www.ivdgl.org.

[7] The EGEE Project, "EGEE Middleware Architecture and Planning (Release 1)," 2004. Available at edms.cern.ch.

[8] P. Eerola *et al.*, "Science on NorduGrid," in *Proceedings of The European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS), 2004*, P. Neittaanmaki and others editors, eds. Jyvaskyla, Finland, July 24^{th}-28^{th}, 2004. Available at `http://www.mit.jyu.fi/eccomas2004/proceedings/pdf/974.pdf`.

[9] R. Pordes *et al.*, "The Open Science Grid," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[10] I. Tomalin *et al.*, "Expected Data Rates from the Silicon Strip Tracker," *CMS NOTE* **2002-047** (2002).

[11] "CMS, The TRIDAS Project, Technical Design Report Volume 2; Data Acquisition and Higher Level Trigger." CERN/LHCC/2002-26, CMS TDR 6.2, 2002.

[12] "LCG 3D project Web Site." Located at `http://lcg3d.cern.ch`.

[13] "Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid." `http://lcg.web.cern.ch/LCG/C-RRB/2005-10/LCG_T0-2_draft_final.pdf`, 2005.

[14] "LHC Computing Grid Technical Design Report." To be released, 2005.

[15] "Central Data Recording (CDR) Web Site." Located at `http://cdr.web.cern.ch/cdr/`.

[16] T. Barrass *et al.*, "Software Agents in Data and Workload Management," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[17] "Storage Resource Management (SRM) project website." Located at `http://sdm.lbl.gov/indexproj.php?ProjectID=SRM`.

[18] L. Lueking, "Access to HEP conditions data using FroNtier: A web-based database delivery system," 2005. Slides available at http://www2.twgrid.org/event/isgc2005.

[19] C. Grandi *et al.*, "Object Based System for Batch Job Submission and Monitoring (BOSS)," *CMS NOTE* **2003-005** (2003).

[20] C. Grandi *et al.*, "BOSS v.4 Architecture," *CMS NOTE (in preparation)* (2005).

[21] "log4cplus at SourceForge." Located at `http://log4cplus.sourceforge.net/`.

[22] A. Nowack, "XCMSI." Web site located at `http://cern.ch/cms-xcmsi`, 2005.

[23] K. Rabbertz *et al.*, "CMS Software Installation," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[24] "CRAB project Web Site." Located at `http://savannah.cern.ch/projects/crab/`.

[25] H. Stockinger and F. Donno, "Data Location Interface for the Workload Management System." Available at `http://cmsdoc.cern.ch/cms/grid/docs/DataLocationInterface.pdf`, 2004.

[26] F. Wuerthwein, "Me - My friends - the grid." Available at http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=128, 2005. OSG-doc-128-v1.

[27] D. Evans, "ShREEK Overview Presentation." Slides available at `http://www.uscms.org/SoftwareComputing/Grid/MCPS/Talks/shreek-overview-pres.pdf`, 2005.

[28] "ShREEK Wiki Pages." Available at `http://lynx.fnal.gov/runjob/ShREEK_20Wiki_20Pages`, 2005.

[29] "Privilege Project Web Site." Located at `http://computing.fnal.gov/docs/products/voprivilege/`.

[30] I. F. et. al., "OSG Authorization Architecture." Available at , 2005.

[31] V. Lefebure and J. Andreeva, "RefDB: The Reference Database for CMS Monte Carlo Production," in *Computing in High-Energy and Nuclear Physics 2003 Conference Proceedings(CHEP 03)*. La Jolla, California, 24-28 March, 2003. `hep-ex/0305092`.

[32] M. Stavrianakou *et al.*, "An Object-Oriented Simulation Program for CMS," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[33] F. Beaudette, "FAMOS, a FAst MOnte-Carlo Simulation for CMS," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[34] M. Stavrianakou *et al.*, "Mantis: a Framework and Toolkit for Geant4-Based Simulation in CMS," *CMS NOTE* **2002-032** (2002).

[35] V. Karimaki *et al.*, "CMKIN v3 User's Guide," *CMS IN* **2004-016** (2004). Available on demand from the CMS secretariat.

[36] N. Neumeister *et al.*, "Muon Reconstruction Software in CMS," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[37] T. Speer *et al.*, "A Gaussian-sum Filter for Vertex Reconstruction," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[38] E. Chabanat *et al.*, "Deterministic Annealing for Vertex Finding at CMS," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[39] S. Cucciarelli *et al.*, "Pixel Reconstruction in the CMS High-Level Trigger," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[40] T. Todorov *et al.*, "Volume-based Representation of the Magnetic Field," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[41] O. van der Aa *et al.*, "High Level Trigger software for the CMS experiment," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[42] G. Zito *et al.*, "Monitoring CMS Tracker Construction and Data Quality Using a Grid/Web Service Based on a Visualization Tool," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo (poster only).

[43] A. Fanfani *et al.*, "Distributed Computing Grid Experiences in CMS DC04," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[44] D. Bonacorsi *et al.*, "Tier-1 and Tier-2 Real-time Analysis Experience in CMS Data Challenge 2004," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[45] N. De Filippis *et al.*, "Tier-1 and Tier-2 Real-time Analysis Experience in CMS DC04," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[46] J. Andreeva *et al.*, "Production Management Software for the CMS Data Challenge," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[47] D. Stickland, "Planning for the 5% Data Challenge, DC04," *CMS IN* **2002-054** (2002).

[48] C. Grandi, "CMS Distributed Data Analysis Challenges," *Nucl. Instrum. Meth.* **A534** (2004) 87–93. ACAT03.

[49] M. Ernst *et al.*, "Distributed Filesystem Evaluation and Deployment at the US-CMS Tier-1 Center," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004.

[50] S. Wynhoff *et al.*, "Using the Reconstruction Software, ORCA, in the CMS Data Challenge 2004," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[51] A. Fanfani *et al.*, "Use of Grid Tools to Support CMS Distributed Analysis," in *Proceedings of the IEEE-NSS'04 Conference*. Rome, Italy, October 16^{th}-22^{nd}, 2004, 2004. to be published.

[52] F. van Lingen *et al.*, "Grid Enabled Analysis: Architecture, Prototype, and Status," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[53] H. Tallini *et al.*, "GROSS: an End User Tool for Carrying Out Batch Analysis of CMS Data on the LCG-2 Grid," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[54] C. Steenberg *et al.*, "The Clarens Grid-enabled Web Services Framework: Services and Implementation," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[55] V. Lefebure *et al.*, "RefDB," *CMS IN* **2002-044** (2002).

[56] G. E. Graham, D. Evans, and I. Bertram, "McRunjob: A High Energy Physics Workflow Planner for Grid Production Processing," in *Computing in High-Energy and Nuclear Physics 2003 Conference Proceedings(CHEP 03)*. La Jolla, California, 24-28 March, 2003. `cs.dc/0305063`.

[57] A. Arbree *et al.*, "Virtual Data in CMS Analysis," in *Computing in High-Energy and Nuclear Physics 2003 Conference Proceedings(CHEP 03)*. La Jolla, California, 24-28 March, 2003. `physics/0306008`.

[58] N. De Filippis *et al.*, "Combined Analysis of GRIDICE and BOSS Information Recorded During CMS-LCG0 Production," *CMS NOTE* **2004-028** (2004).

[59] D. Bonacorsi *et al.*, "Running CMS Software on Grid Testbeds," in *Computing in High-Energy and Nuclear Physics 2003 Conference Proceedings(CHEP 03)*. La Jolla, California, 24-28 March, 2003. `physics/0306038`.

[60] A. Osman *et al.*, "Resource Monitoring Tool for CMS production," *CMS NOTE* **2003-013** (2003).

[61] T. Wildish *et al.*, "The Spring 2002 DAQ TDR Production," *CMS NOTE* **2002-034** (2002).

[62] P. Capiluppi *et al.*, "CMS Test of the European DataGrid Testbed," *CMS NOTE* **2003-014** (2003).

[63] S. Dasu *et al.*, "Use of Condor and GLOW for CMS Simulation Production," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[64] S. Schmid, "Study and Prototype Implementation of a Distributed System," *CMS NOTE* **2004-010** (2004).

[65] R. Harris *et al.*, "Report of the CMS Data Management RTAG," *CMS IN* **2004-038** (2004). Available on demand from the CMS secretariat.

[66] I. Vorobiev *et al.*, "Status and Perspectives of Detector Databases in the CMS Experiment at the LHC," *CMS NOTE* **2004-026** (2004).

[67] Z. Xie *et al.*, "Modeling a Hierarchical Data Registry with Relational Databases in a Distributed Environment," *CMS IN* **2004-025** (2004). Available on demand from the CMS secretariat.

[68] A. Aerts, M. Liendl, and R. Gomez-Reino, "Detector Geometry Database," *CMS IN* **2004-011** (2004). Available on demand from the CMS secretariat.

[69] A. J. Muhammad *et al.*, "Migration of the XML Detector Description Data and Schema to a Relational Database," *CMS NOTE* **2003-031** (2003).

[70] R. McClatchey *et al.*, "Deserializing Object Data while Schemas Evolve," *CMS NOTE* **2002-029** (2002).

[71] Z. Xie and V. Innocente, "Evaluation of Oracle9i C++ Call Interface," *CMS NOTE* **2002-012** (2002).

[72] I. Osborne and G. Raymond, "3D Graphics Under Linux," *CMS IN* **2002-041** (2002).

[73] I. Osborne, L. Taylor, and L. A. Tuura, "IGUANA Plan for 2002," *CMS IN* **2002-018** (2002).

[74] S. Iqbal, "Evaluation of Oracle9i to Manage CMS Event Store: Oracle Architecture to Store Petabyte of Data (Part One)," *CMS IN* **2002-002** (2002).

[75] I. Osborne *et al.*, "Composite Framework for CMS User Applications," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[76] M. Stavrianakou *et al.*, "Mantis: the Geant4-based Simulation Specialization of the CMS COBRA Framework," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[77] M. Case *et al.*, "CMS Detector Description: New Developments," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[78] A. Aerts *et al.*, "A Database Perspective on CMS Data," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[79] R. Brun and F. Rademakers, "ROOT - An Object Oriented Data Analysis Framework," in *AIHENP'96 Workshop*, volume Phys. Res. A 389, pp. 81–86. Lausanne, Switzerland, September, 1996, 1997.

[80] C. Grandi, "Use Cases and Requirements for Software Installation in Grid and End-User Desktop Environments," *CMS IN* **2004-047** (2004). Available on demand from the CMS secretariat.

[81] D. Chamont and C. Charlot, "OVAL: the CMS Testing Robot," in *Computing in High-Energy and Nuclear Physics 2003 Conference Proceedings(CHEP 03)*. La Jolla, California, 24-28 March, 2003. `cs.se/0306054`.

[82] G. Eulisse, "Installation/Usage Notes For Oprofile," *CMS IN* **2002-053** (2002).

[83] L. Taylor, "CMS Software Quality," *CMS IN* **2002-050** (2002).

[84] I. Osborne *et al.*, "Evaluation Of the CMT and SCRAM Software Configuration, Build And Release Management Tools," *CMS IN* **2002-046** (2002).

[85] S. Schmid *et al.*, "Parallel Compilation of CMS Software," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[86] H. Wellish *et al.*, "PRS Software Quality Policy," *CMS IN* **2002-037** (2002).

[87] G. Thomson, "Software Metrics Report of CMS Reconstruction Software," *CMS IN* **2002-033** (2002).

[88] L. Taylor, "CMS Computing and Software Tasks and Manpower for 2003-2007," *CMS IN* **2003-038** (2003). Available on demand from the CMS secretariat.

[89] V. Innocente, L. Taylor, and D. Stickland, "Computing And Core Software (CCS) Schedule And Milestones: Version 33," *CMS IN* **2002-039** (2002).

[90] D. Stickland, "Planning for the Computing TDR," *CMS IN* **2002-059** (2002).

[91] D. Stickland *et al.*, "Proposed Scope And Organization Of CMS-CPT. Computing And Core Software, Physics Reconstruction and Selection, TriDAS (Online Computing)," *CMS IN* **2002-038** (2002).

[92] C. Grandi *et al.*, "CMS Grid Implementation Plan - 2002," *CMS NOTE* **2002-015** (2002).

[93] C. Grandi (for the CMS Computing and Core Software Group), "Plans for the Integration of Grid Tools in the CMS Computing Environment," in *Computing in High-Energy and Nuclear Physics 2003 Conference Proceedings(CHEP 03)*. La Jolla, California, 24-28 March, 2003. `hep-ex/0305099`.

[94] P. Sphicas *et al.*, "PRS Acquisition Computing Software: Scope and Organization of CMS-CPT," *CMS IN* **2002-068** (2002).

[95] M. Burgon-Lyon *et al.*, "JIM Deployment for the CDF Experiment," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[96] I. Sfiligoi *et al.*, "Monitoring the CDF Distributed Computing Farms," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo (slides only).

[97] V. Bartsch *et al.*, "Testing the CDF Distributed Computing Framework," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[98] A. Kreymer *et al.*, "Tools for GRID Deployment of CDF Offline and SAM Data Handling Systems for Summer 2004 Computing," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[99] A. Sill *et al.*, "Globally Distributed User Analysis Computing at CDF," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[100] S. Stonjek *et al.*, "Deployment of SAM for the CDF Experiment," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[101] I. Sfiligoi *et al.*, "The Condor based CDF CAF," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[102] B. Quinn *et al.*, "Performance of an operating High Energy Physics Data grid, DØSAR-grid," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004.

[103] T. Harenberg *et al.*, "DØ Data Processing within EDG/LCG," in *Proceedings of the CHEP'04 Conference*. Interlaken, Switzerland, September 27^{th} - October 1^{st}, 2004. Published on InDiCo.

[104] G. Garzoglio *et al.*, "Experience Using Grid Tools for CDF Physics," in *Proceedings of the ACAT'03 Conference*, S. Kawabata and D. Perret-Gallix, eds., volume A 534, pp. 38–41. Tsukuba, Japan, December 1^{st}-5^{nd}, 2003, 2004.