GRID технологии.
Или как CERN си прави сметките.

Тодор Иванов

Септември 08, 2017

Ускорителя LHC

- HEP + IT = Проблем.

Ускорителя LHC

- HEP + IT = Проблем.

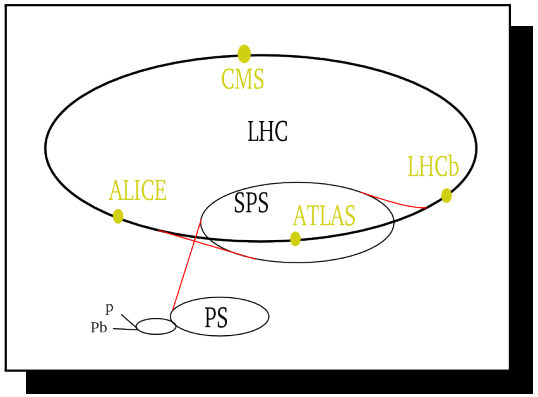
- Максимална светимост 2017:

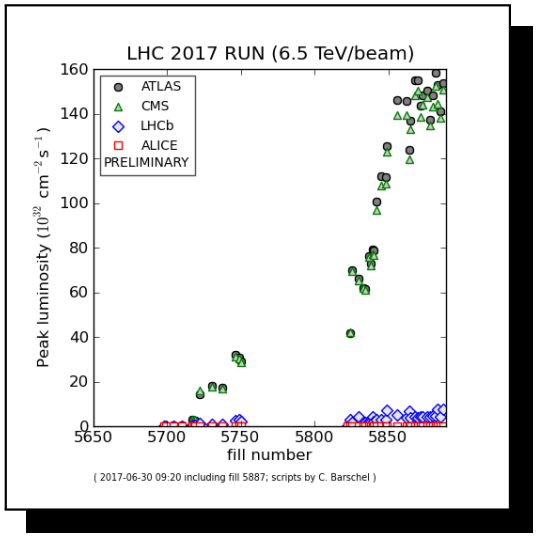
$$1.58 * 10^{34} [cm^{-2}.s^{-1}]$$

Отстояние между групите в снопа: 25 ns
Честота на сблъсъците: 40 MHz

$$N_{events} = L\sigma_{event}$$

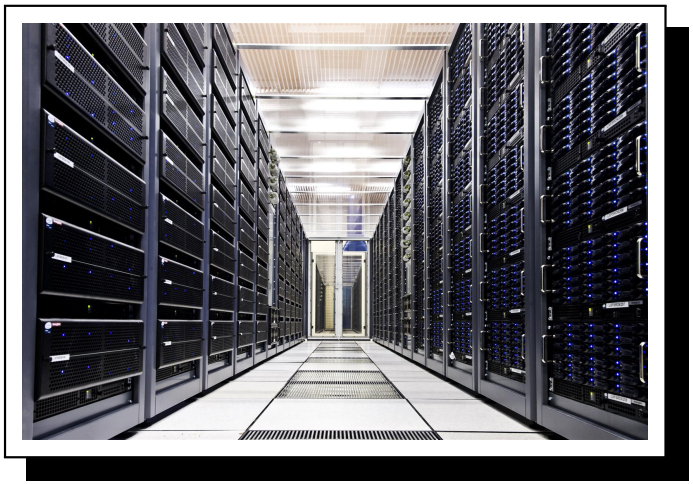
$\sigma_{p-p/7Tev}$ $60mbarn$
600 000 000 000 сблъсаци / сек
Размер на данните: 6 Gbyte/s





CERN - Изчислителен център

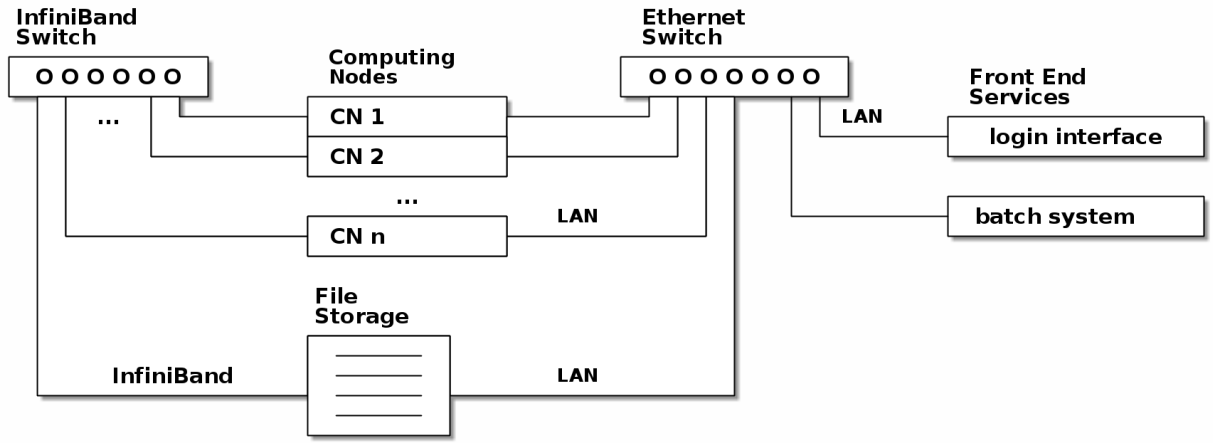
$Meyrin + Vigner:$	
$Memory(TiB)$	1153.72
$Numberofcores$	217530
$Numberofdisks$	90179
$Numberofmemorymodules$	108743
$Numberofprocessors$	27409
$Numberofsystems$	14421
$RawHDDcapacity(TiB)$	246262



source: http://hwcollect.cern.ch/

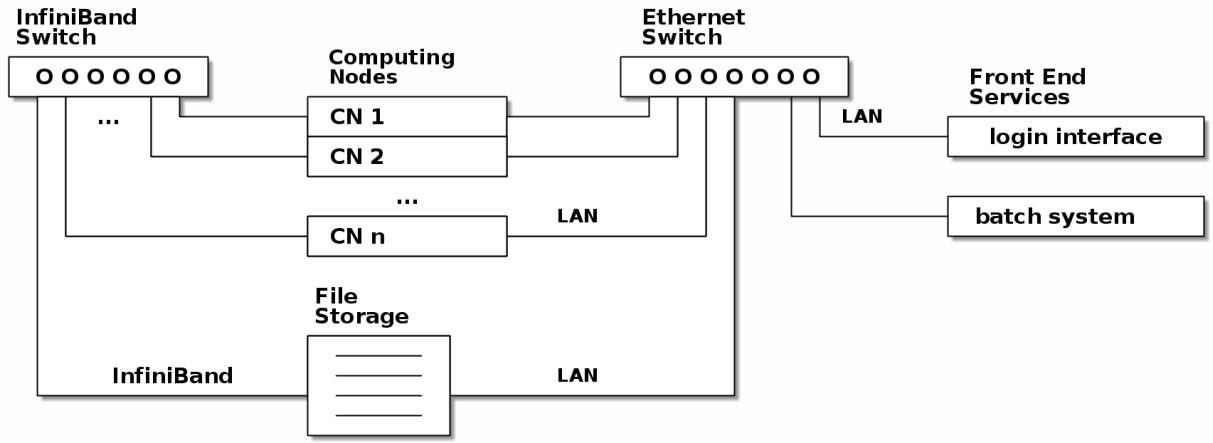
Grid: Дефиниция

- Стандартни клъстърни системи



Grid: Дефиниция

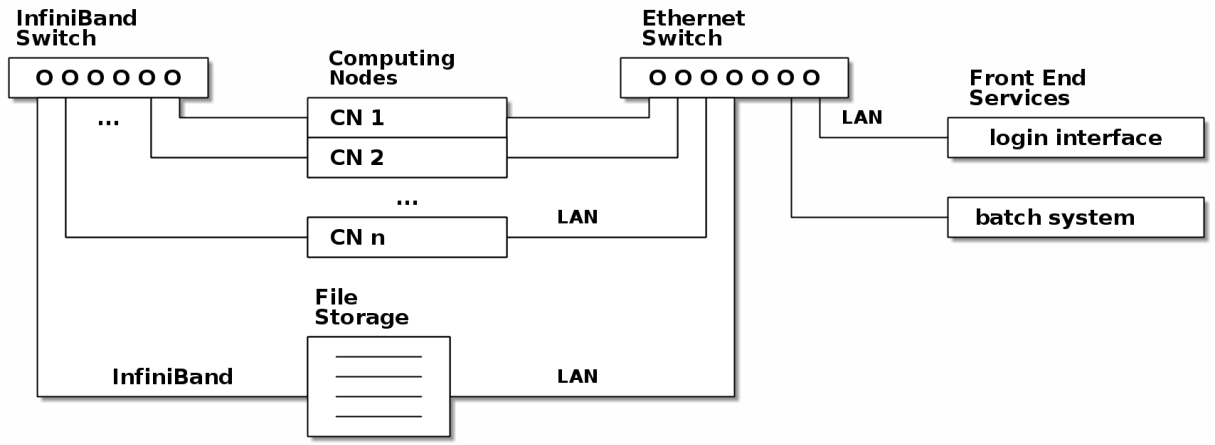
- Стандартни клъстърни системи



- Обединяване в мрежа от клъстърни системи:

Grid: Дефиниция

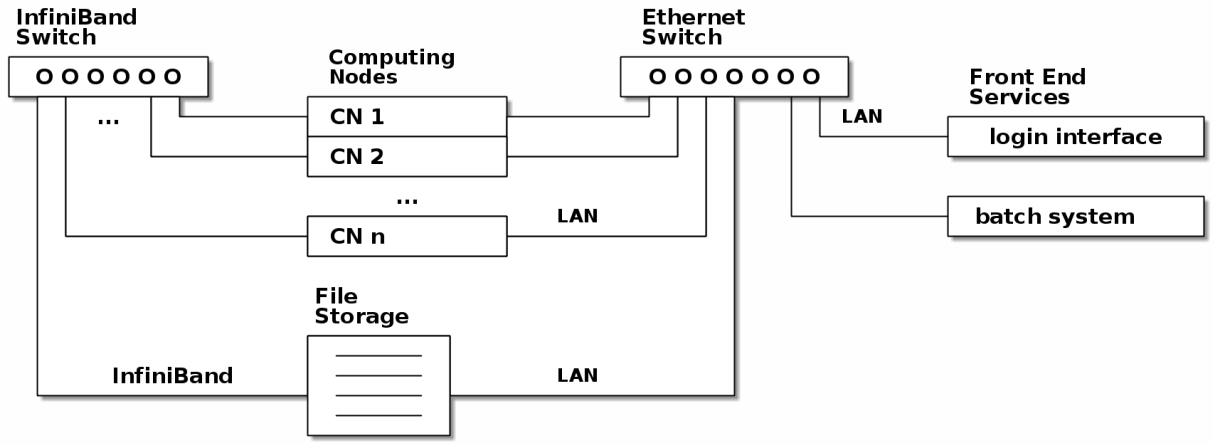
- Стандартни клъстърни системи



- Обединяване в мрежа от клъстърни системи:
- Нива на абстракция

Grid: Дефиниция

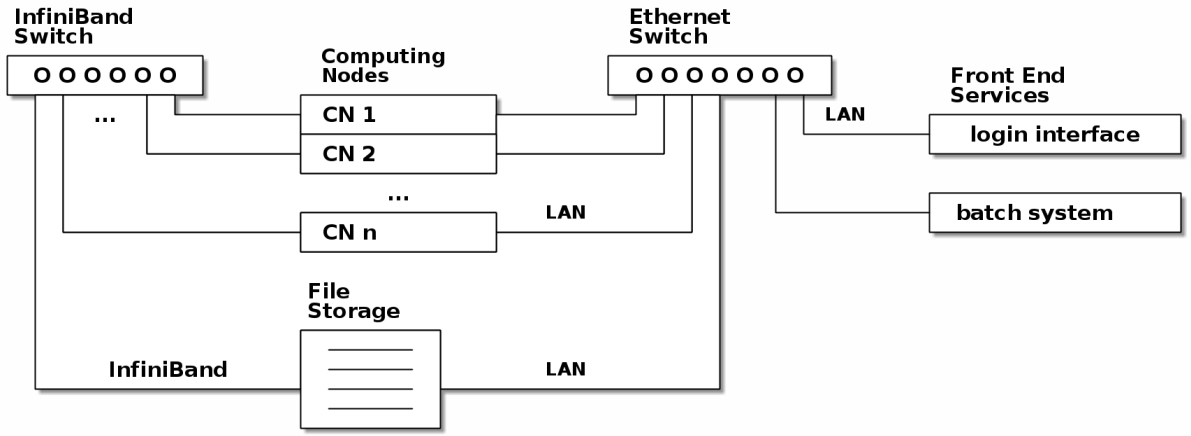
- Стандартни клъстърни системи



- Обединяване в мрежа от клъстърни системи:
- Нива на абстракция

Grid: Дефиниция

- Стандартни клъстърни системи



- Обединяване в мрежа от клъстърни системи:
- Нива на абстракция

Grid: Дефиниция

Стъпки за обединение на ниво услуги:

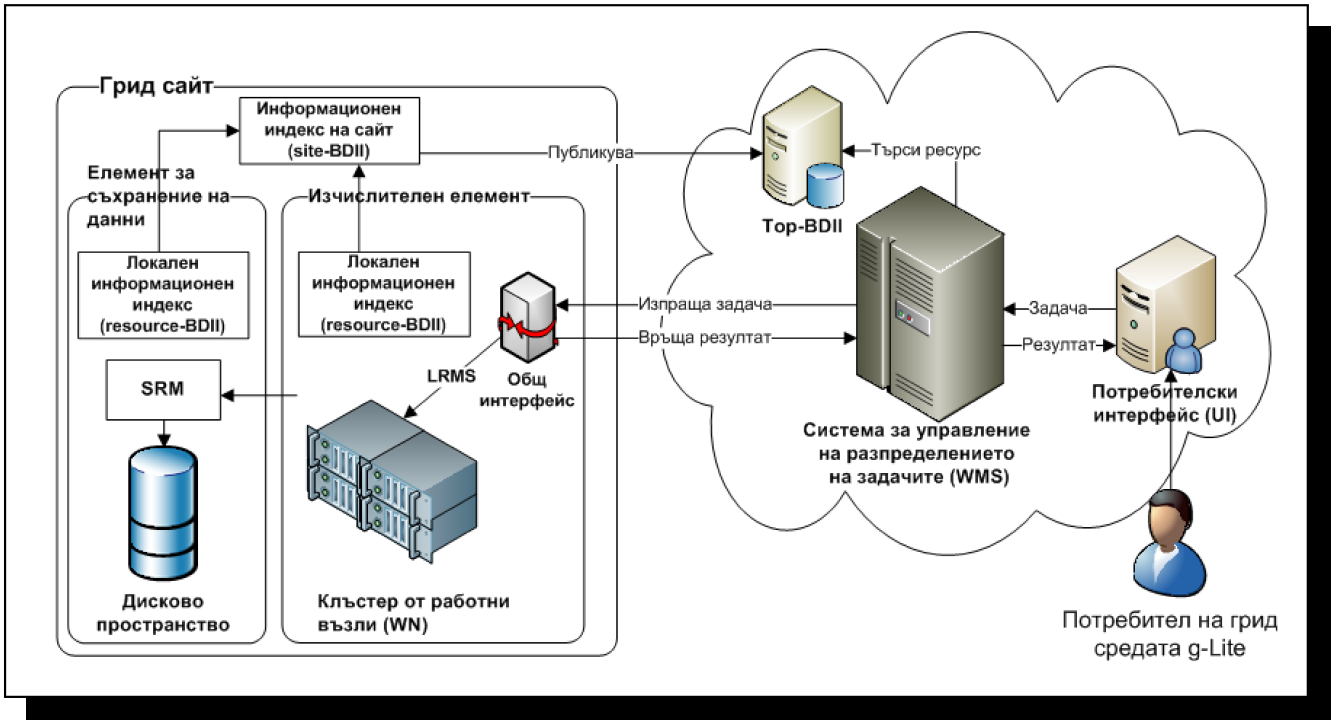
- Базирани на P2P комуникация - силно децентрализирани.
  - Napster - 1999г.
  - Seti@Home - 1999г. University of California, Berkeley
- Усложняване на видовете услуги и връзките между тях:

- Мидълуеър: gLite - част от проекта EGEE/EMI; Globustoolkit; VDT (Virtual Data Toolkit).

Grid: Дефиниция

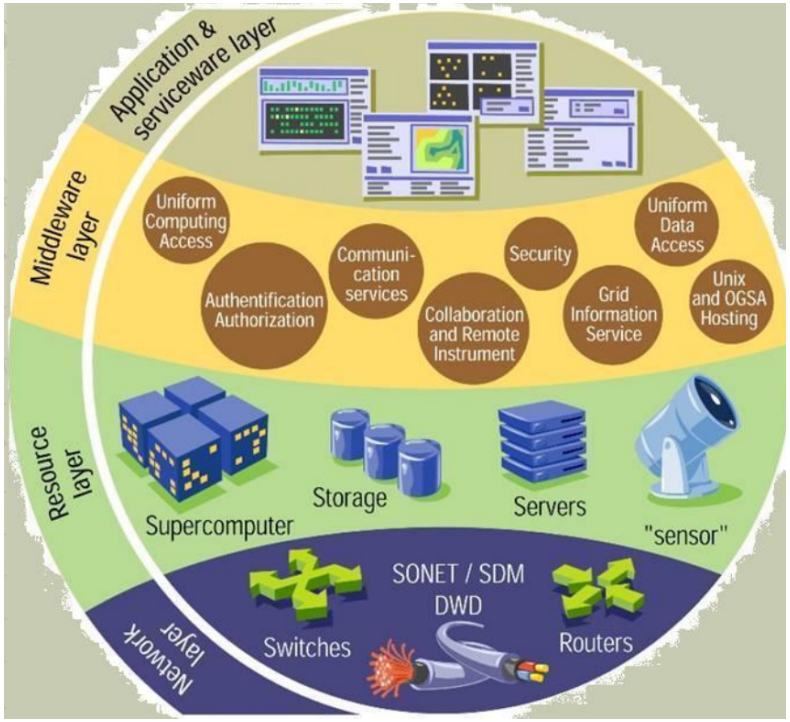
Стъпки за обединение на ниво услуги:

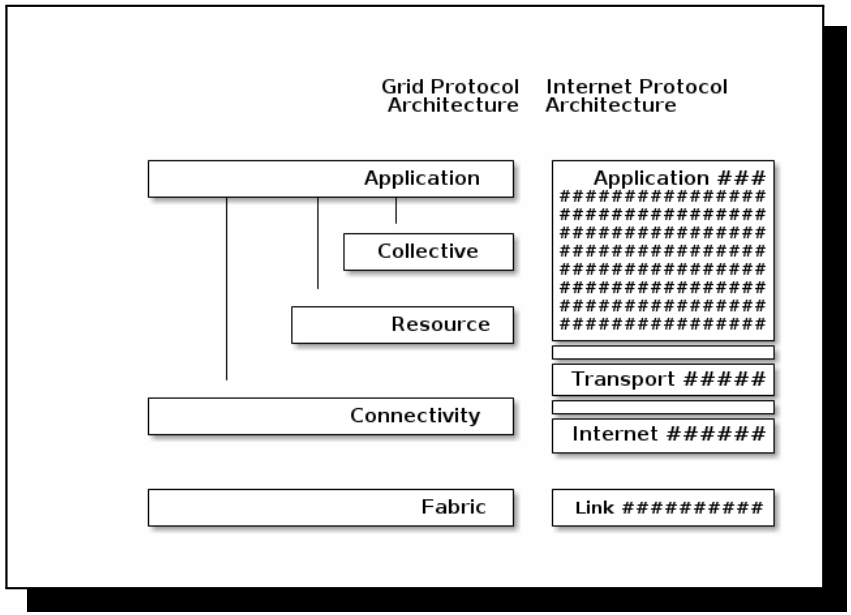
- Базирани на P2P комуникация - силно децентрализирани.
  - Napster - 1999г.
  - Seti@Home - 1999г. University of California, Berkeley
- Усложняване на видовете услуги и връзките между тях:



- Мидълуеър: gLite - част от проекта EGEE/EMI; Globustoolkit; VDT (Virtual Data Toolkit).

Грид - Слоеве


Application & serviceware layer
Middleware layer
Resource layer
Network layer
Uniform Computing Access
Authentification Authorization
Communi-cation services
Security
Collaboration and Remote Instrument
Grid Information Service
Uniform Data Access
Unix and OGSA Hosting
Supercomputer
Storage
Servers
"sensor"
SONET / SDM
DWD
Switches
Routers


Grid Protocol Architecture
Internet Protocol Architecture
Application
Collective
Resource
Connectivity
Fabric
Application ###
Transport #####
Internet ######
Link ##########

Grid: Дефиниция

- Виртуални организации.
- Изисквания към една мрежа от ресурси за да бъде Грид:
  1. Да включва в себе си и да координира споделянето на някакви географски разделени ресурси и потребители.
  2. Да използва стандартни, отворени протоколи и интерфейси с общо предназначение.
  3. Да предоставя достатъчно надежден метод за осигуряване и оценка на качеството на специализирани (QoS) ресурси.
- деф: Мрежа от ресурси с теоретично неограничен капацитет, при която основният модел на управление на ресурсите, това е принципът на “предоставяне при поискване” и притежаваща възможността за виртуализация и споделяне на самите ресурси.

Grid: Дефиниция

- Виртуални организации.
- Изисквания към една мрежа от ресурси за да бъде Грид:
  1. Да включва в себе си и да координира споделянето на някакви географски разделени ресурси и потребители.
  2. Да използва стандартни, отворени протоколи и интерфейси с общо предназначение.
  3. Да предоставя достатъчно надежден метод за осигуряване и оценка на качеството на специализирани (QoS) ресурси.
- деф: Мрежа от ресурси с теоретично неограничен капацитет, при която основният модел на управление на ресурсите, това е принципът на “предоставяне при поискване” и притежаваща възможността за виртуализация и споделяне на самите ресурси.
- Примери:
  Particle Physics Data Grid (PPDG); EU-Datagrid ; NASA’s Information Power Grid;

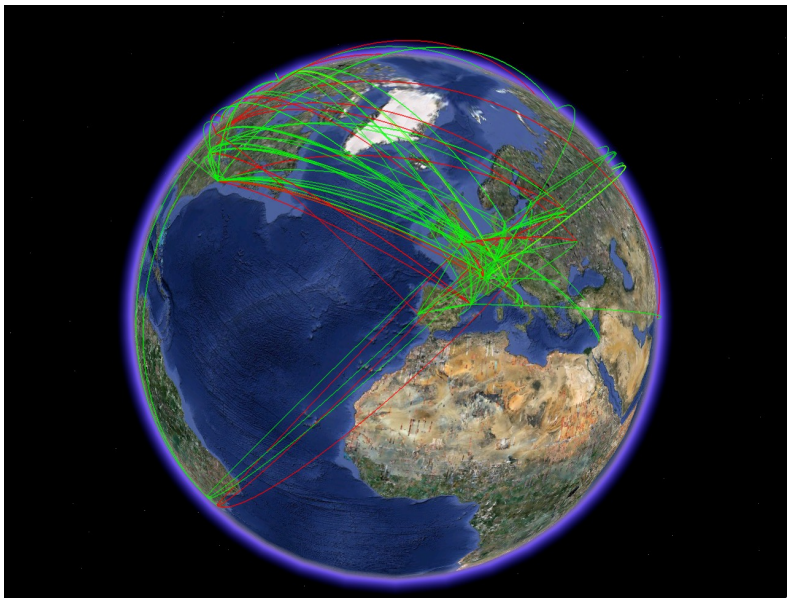
Grid: Дефиниция

- Виртуални организации.
- Изисквания към една мрежа от ресурси за да бъде Грид:
  1. Да включва в себе си и да координира споделянето на някакви географски разделени ресурси и потребители.
  2. Да използва стандартни, отворени протоколи и интерфейси с общо предназначение.
  3. Да предоставя достатъчно надежден метод за осигуряване и оценка на качеството на специализирани (QoS) ресурси.
- деф: Мрежа от ресурси с теоретично неограничен капацитет, при която основният модел на управление на ресурсите, това е принципът на “предоставяне при поискване” и притежаваща възможността за виртуализация и споделяне на самите ресурси.
- Примери:
  Particle Physics Data Grid (PPDG); EU-Datagrid ; NASA’s Information Power Grid;

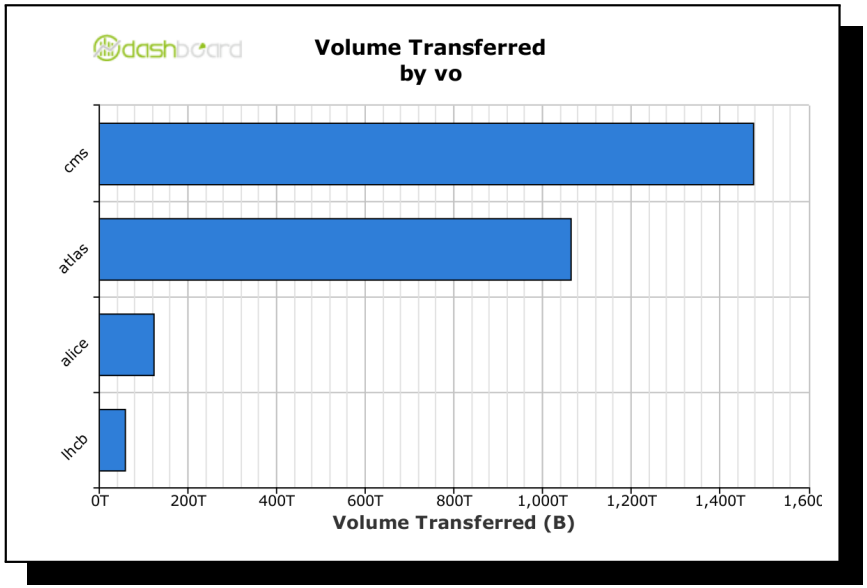
Участници в WLCG

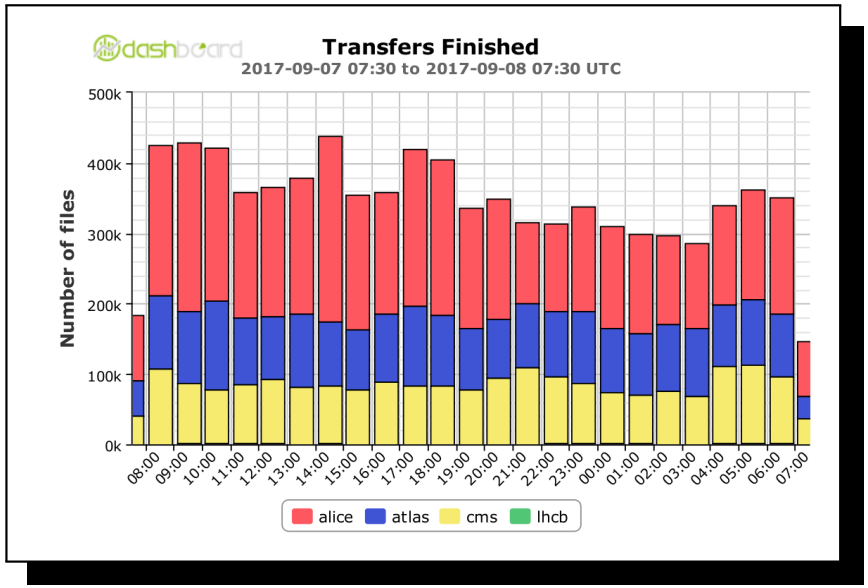
WLCG
Worldwide LHC Computing Grid

WLCG = EGI + OSG + NORDU GRID
– 40 държави
– 170 изчислителни центъра
– 2 милиона изчислителни задачи/ден
– 10 Gb/s глобална скорост
– Над 10 000 потребителя

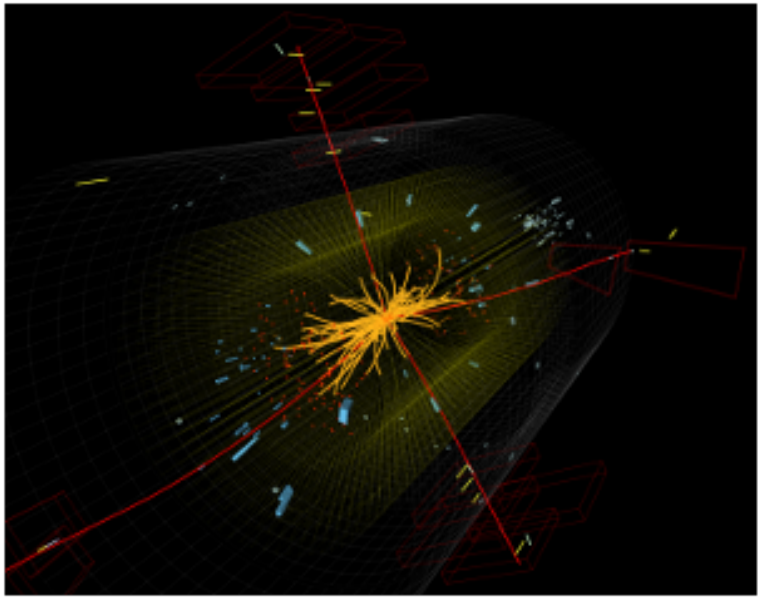


Капацитет на WLCG

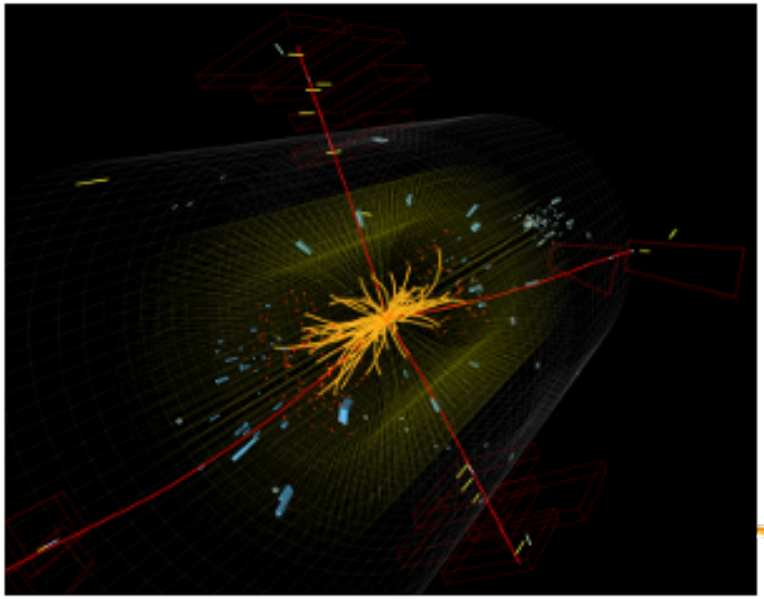




Абстрактен модел на данните.

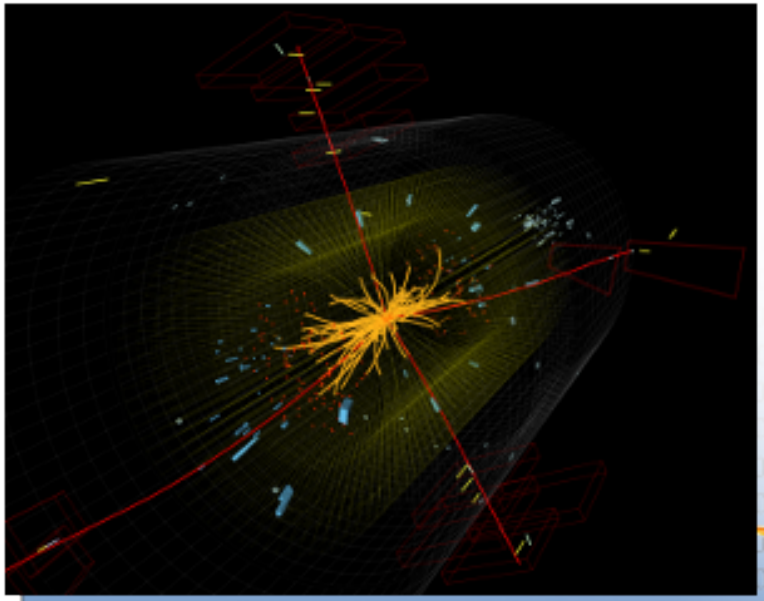


Абстрактен модел на данните.



$$S[\mathbf{q}(\mathbf{t})] = \int_{t_1}^{t_2} L(\mathbf{q}, \dot{\mathbf{q}}, t)dt$$

Абстрактен модел на данните.



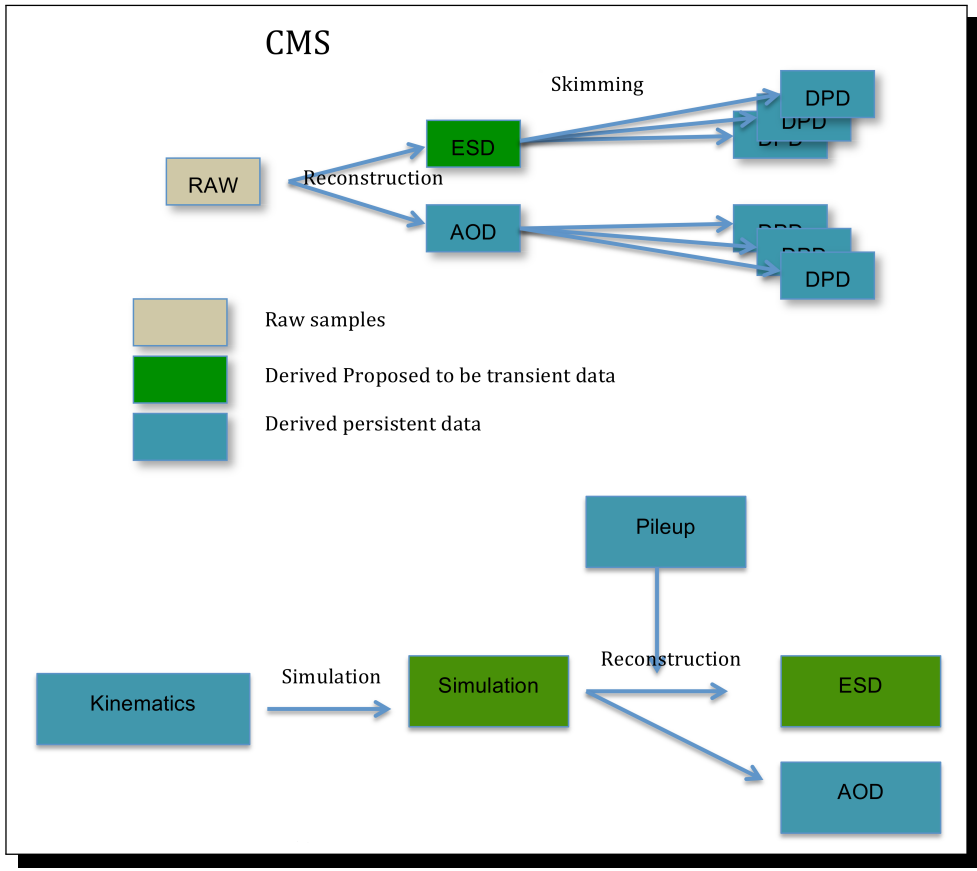
$$S[\mathbf{q}(\mathbf{t})] = \int_{t_1}^{t_2} L(\mathbf{q}, \dot{\mathbf{q}}, t)dt$$

0x01e84cf0:	0x01e8	0x87ec	0x01e8	0x85d8	0x7363	0x616e	0x0000	0x0000
0x01e84d00:	0x0000	0x0019	0x0000	0x0000	0x0000	0x0000	0x01e8	0x5b7c
0x01e84d10:	0x01e8	0x87e8	0x01e8	0x8618	0x7365	0x7400	0x0000	0x0000
0x01e84d20:	0x0000	0x0019	0x0000	0x0000	0x0000	0x0000	0x01e8	0x5b7c
0x01e84d30:	0x01e8	0x87a8	0x01e8	0x8658	0x7370	0x6c69	0x7400	0x0000
0x01e84d40:	0x0000	0x0019	0x0000	0x0000	0x0000	0x0000	0x01e8	0x5b7c
0x01e84d50:	0x01e8	0x8854	0x01e8	0x8698	0x7374	0x7269	0x6e67	0x0000
0x01e84d60:	0x0000	0x0019	0x0000	0x0000	0x0000	0x0000	0x01e8	0x5b7c
0x01e84d70:	0x01e8	0x875c	0x01e8	0x86d8	0x7375	0x6273	0x7400	0x0000
0x01e84d80:	0x0000	0x0019	0x0000	0x0000	0x0000	0x0000	0x01e8	0x5b7c
0x01e84d90:	0x01e8	0x87c0	0x01e8	0x8718	0x7377	0x6974	0x6368	0x0000

- Идентификатор на събитие
- Данни от детектора

Тиер структура на данните

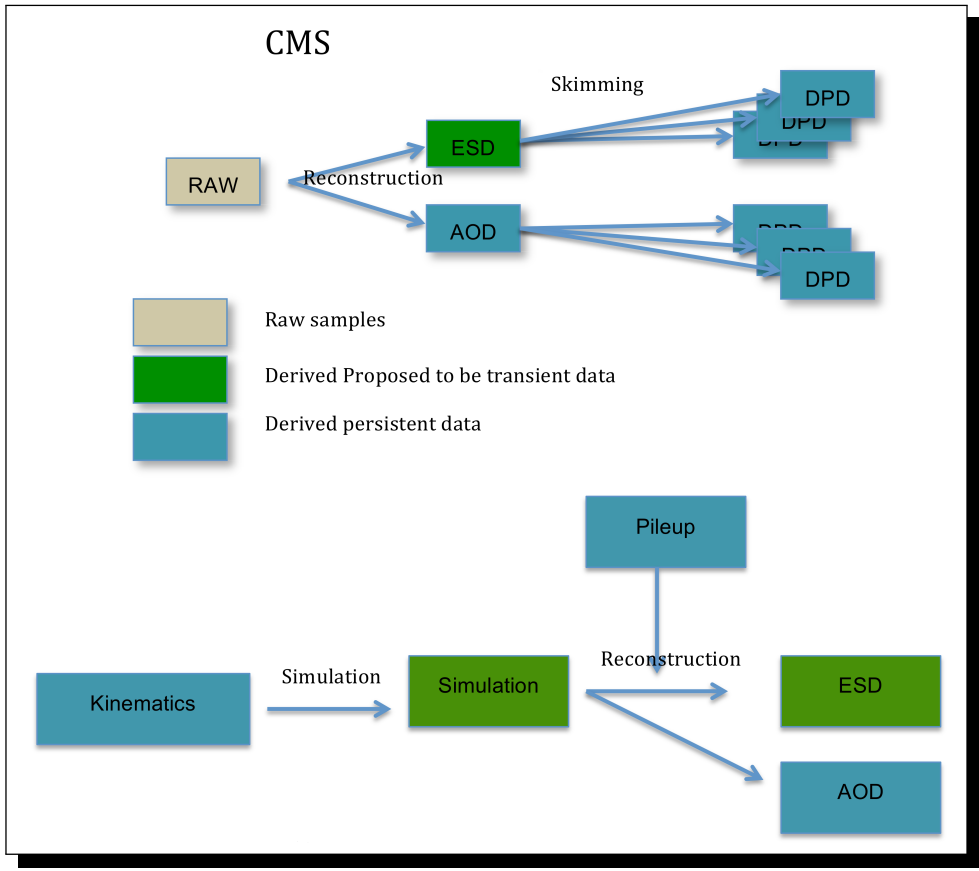
Тясно свързана с работните потоци.



Формат	Съдържание	Размер/съб. [KB]
Типове данни от реални събития.		
DAQ-RAW	Първични - FED формат + L1 тригера	150
RAW	Първични - след онлайн форматиране - от L1 тригер и от HLT	500
ESD (RECO)	Реконструирани - трекове, вертекси, струи, частици + информацията от RAW	1000
AOD	Реконструирани - реконструираните физични обекти.	300
DPD	Извлечени Физични Данни - подбрани AOD данни, специфични за някакъв конкретен физичен анализ	tmp
NTUP	Свързани Списъци	tmp
Типове данни от симулирани събития.		
HEPMC	Изход от генератор на събития	1000
HITS	Симулирани събития - симулираната енергия, оставена в детектора от генерираните събития	tmp
RDO	Симулирани Първични данни - и (RAW-SIM-RECO)	1500

Тиер структура на данните

Тясно свързана с работните потоци.



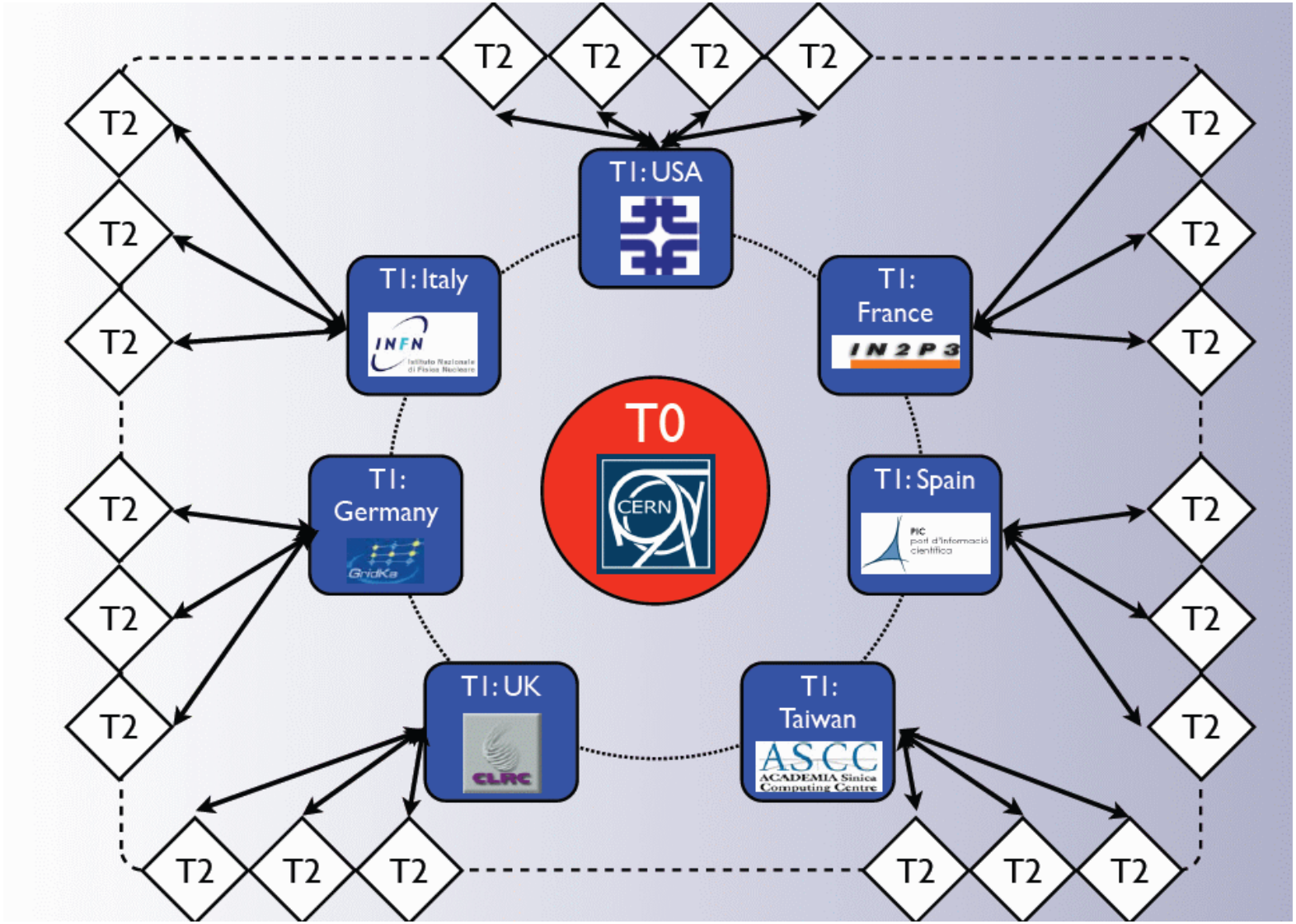
Формат	Съдържание	Размер/съб. [KB]
Типове данни от реални събития.		
DAQ-RAW	Първични - FED формат + L1 тригера	150
RAW	Първични - след онлайн форматиране - от L1 тригер и от HLT	500
ESD (RECO)	Реконструирани - трекове, вертекси, струи, частици + информацията от RAW	1000
AOD	Реконструирани - реконструираните физични обекти.	300
DPD	Извлечени Физични Данни - подбрани AOD данни, специфични за някакъв конкретен физичен анализ	tmp
NTUP	Свързани Списъци	tmp
Типове данни от симулирани събития.		
HEPMC	Изход от генератор на събития	1000
HITS	Симулирани събития - симулираната енергия, оставена в детектора от генерираните събития	tmp
RDO	Симулирани Първични данни - и (RAW-SIM-RECO)	1500

Година	Честота	Изход
2012	460 Hz	328 MB/s
2015	1000 Hz	600 MB/s

Таблица: Честота на работа на тригера от високо ниво CMS.

Тиер структура на изчислителния модел на CMS.

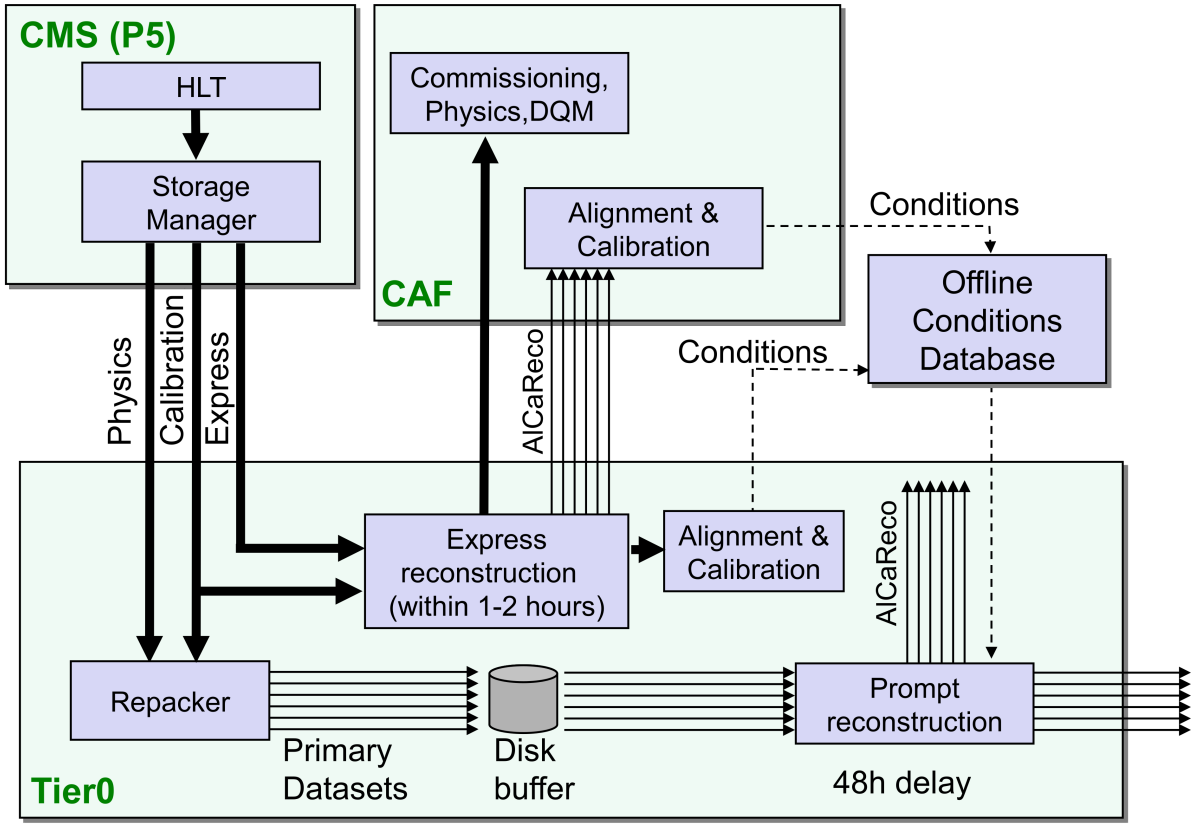
Архитектурния модел е пряко свързан с абстрактния модел на данните.



Тиер0

3 потока от данни

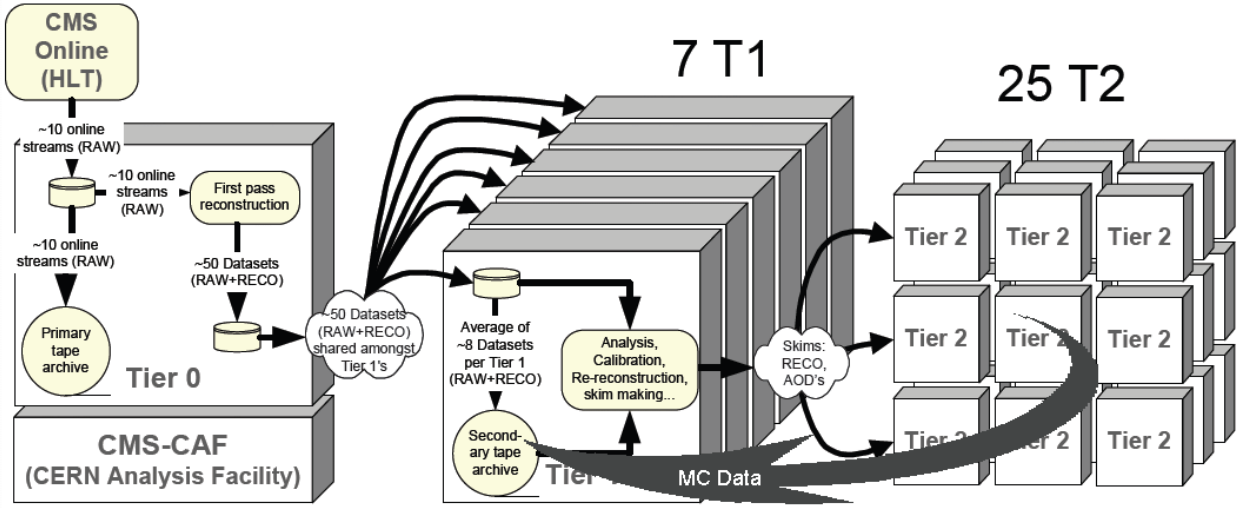
- Физичен поток от данни с честота 200 Hz.
- Експресен поток от данни с честота 20 Hz.
- Поток за калибриране и наблюдение 20 Hz.



5 работни потока

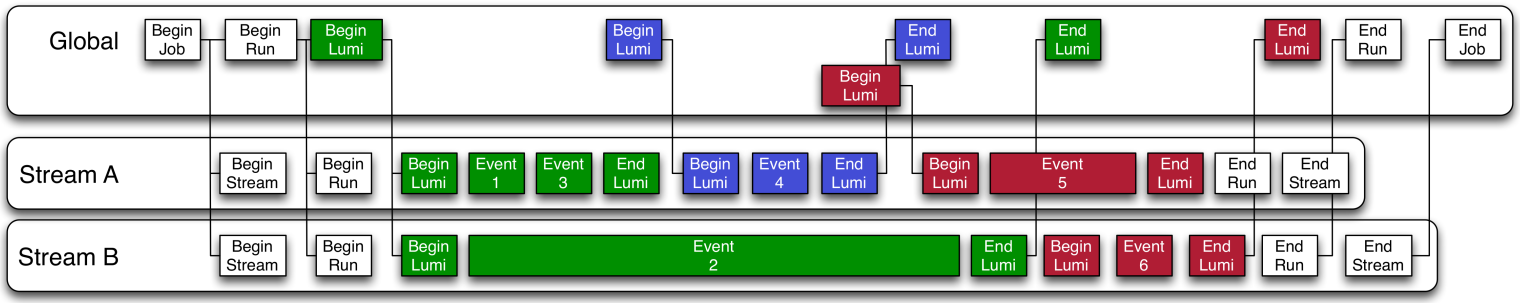
- Сливане
- Препакетиране
- Бърза реконструкция
- Експресна обработка
- Бърза калибровка

от Тиер0 до Тиер2



Тип на изчислителните задачи:

- CPU интензивни
- Груба гранулираност
- Дълго време на изчисления
- Липса на комуникация



Развитие на изчислителеления модел на CMS

asdf

БЛАГОДАРЯ ЗА ВНИМАНИЕТО!

Характеристикина клъстъра BG05-SUGrid

Сайт от ниво Тиер3.

40 x COMPUTING NODES:		8 x STORAGE:	
CPUs:	80 Intel Xeon E5345	RAID arrays:	6
Cores:	160	Total:	72.8 TiB
Threads:	320	Configured Arrays:	RAID6
RAM:	640 GB	Distributed Filesystem:	dCache
Lan:	1 GigE		

Поддържани Виртуални Организации

- CMS - Физика на високите енергии.
- Padme - Физика на високите енергии.
- Biomed - Био-инженерни технологии, Биология, Медицина.
- Envir - Българска ВО - моделирането и защита на околната среда.
- Dteam - Изпълнява мониториращa роля.
- Ops - Изпълнява мониториращa роля.

Производителност на BG05-SUGrid

Теста HEP-SPEC06 - оптимизиран вариант на SPECCPU2006 (all-cp).

CPU	HEP06	Clock (MHz)	L2+L3 (KB)	Core	Memory (GB)	Mainboard
Intel Xeon E5345	58.98	2333	16384	8	16 = (8x2 FB-DDR2-677)	Supermicro X7DBE

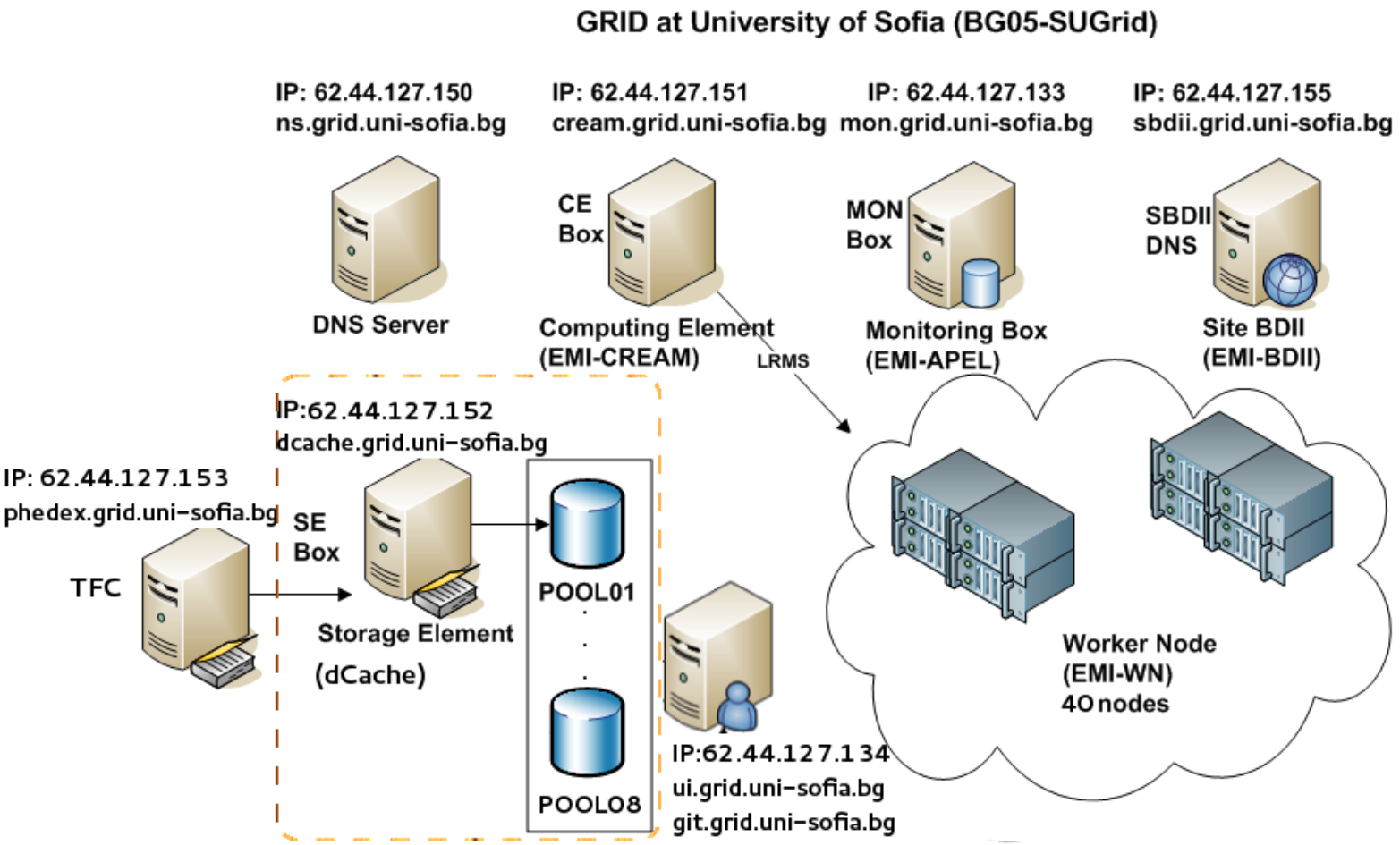
Таблица: Измерване на производителността на един изчислителен възел идентичен с тези от BG05-SUGrid .

$$HEP-SPEC06(BG05-SUGrid\ ) = 40 * 58.98 = 2352 \quad (1)$$

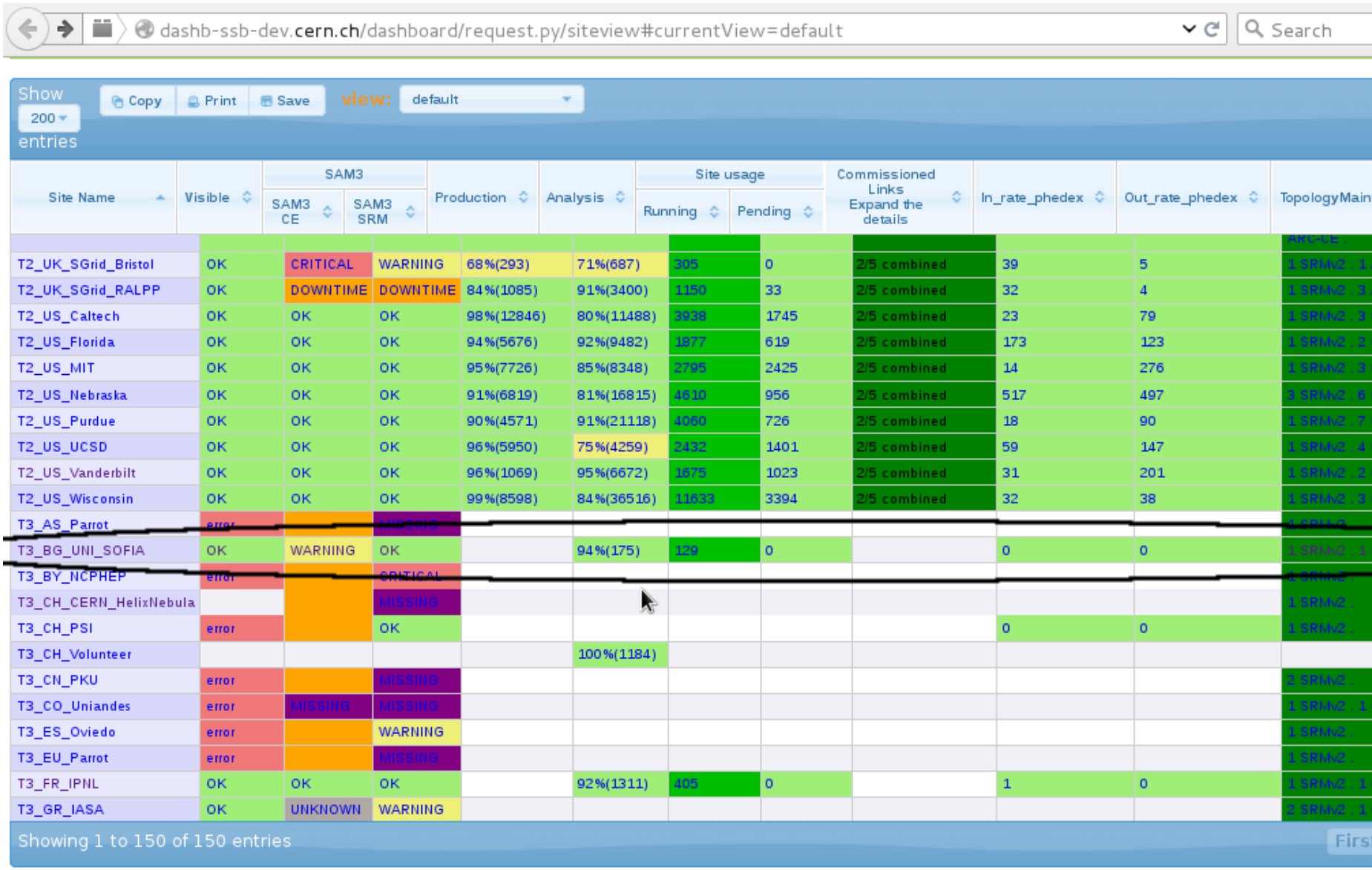
Теоретична паралелна ефективност:

$$\frac{f * N * s}{1000} = (2333 * 160 * 4)/1000 = 1493.12\ [Gflops] \quad (2)$$

Топология на клъстъра BG05-SUGrid



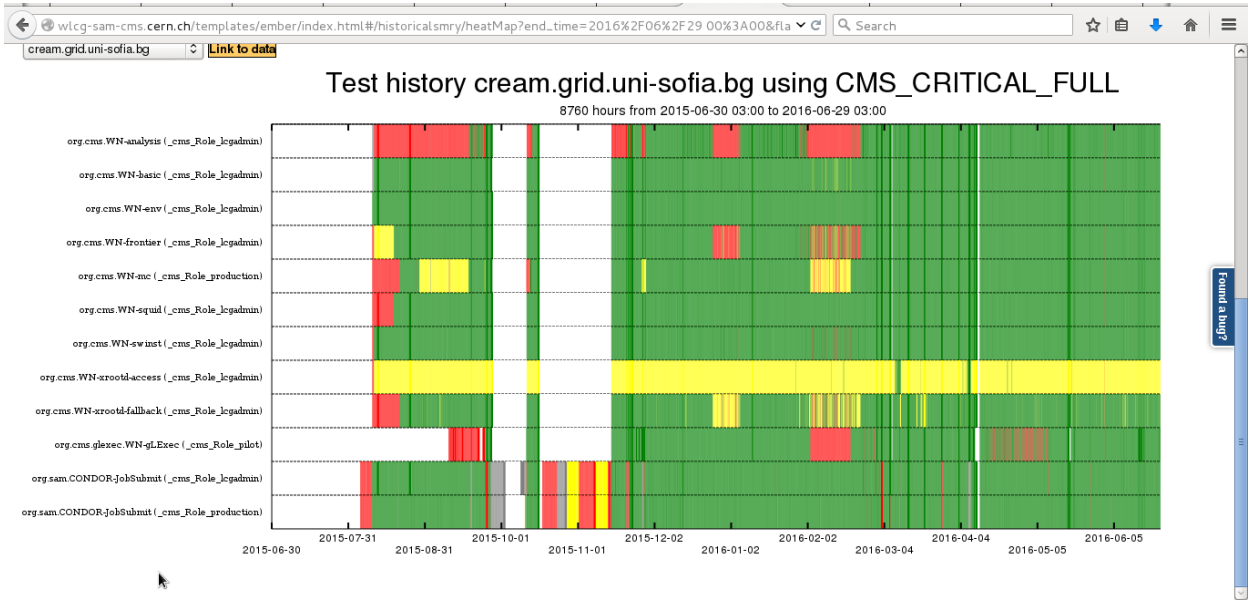
T3_BG_UNI_SOFIA



Site Name	Visible	SAM3		Production	Analysis	Site usage		Commissioned Links Expand the details	In_rate_phedex	Out_rate_phedex	TopologyMain
		SAM3 CE	SAM3 SRM			Running	Pending				
T2_UK_SGrid_Bristol	OK	CRITICAL	WARNING	68%(293)	71%(687)	305	0	2/5 combined	39	5	1 SRMv2 1
T2_UK_SGrid_RALPP	OK	DOWNTIME	DOWNTIME	84%(1085)	91%(3400)	1150	33	2/5 combined	32	4	1 SRMv2 3
T2_US_Caltech	OK	OK	OK	98%(12846)	80%(11488)	3938	1745	2/5 combined	23	79	1 SRMv2 3
T2_US_Florida	OK	OK	OK	94%(5676)	92%(9482)	1877	619	2/5 combined	173	123	1 SRMv2 2
T2_US_MIT	OK	OK	OK	95%(7726)	85%(8348)	2795	2425	2/5 combined	14	276	1 SRMv2 3
T2_US_Nebraska	OK	OK	OK	91%(6819)	81%(16815)	4610	956	2/5 combined	517	497	3 SRMv2 6
T2_US_Purdue	OK	OK	OK	90%(4571)	91%(21118)	4060	726	2/5 combined	18	90	1 SRMv2 7
T2_US_UCSD	OK	OK	OK	96%(5950)	75%(4259)	2432	1401	2/5 combined	59	147	1 SRMv2 4
T2_US_Vanderbilt	OK	OK	OK	96%(1069)	95%(6672)	1675	1023	2/5 combined	31	201	1 SRMv2 2
T2_US_Wisconsin	OK	OK	OK	99%(8598)	84%(36516)	11633	3394	2/5 combined	32	38	1 SRMv2 3
T3_AS_Parrot	error		MISSING								1 SRMv2
T3_BG_UNI_SOFIA	OK	WARNING	OK		94%(175)	129	0		0	0	1 SRMv2 1
T3_BY_NCPHEP	error		CRITICAL								1 SRMv2
T3_CH_CERN_HelixNebula			MISSING								1 SRMv2
T3_CH_PSI	error		OK						0	0	1 SRMv2
T3_CH_Volunteer					100%(1184)						
T3_CN_PKU	error		MISSING								2 SRMv2
T3_CO_Uniandes	error	MISSING	MISSING								1 SRMv2 1
T3_ES_Oviedo	error		WARNING								1 SRMv2
T3_EU_Parrot	error		MISSING								1 SRMv2
T3_FR_IPNL	OK	OK	OK		92%(1311)	405	0		1	0	1 SRMv2 1
T3_GR_IASA	OK	UNKNOWN	WARNING								2 SRMv2 1

Showing 1 to 150 of 150 entries

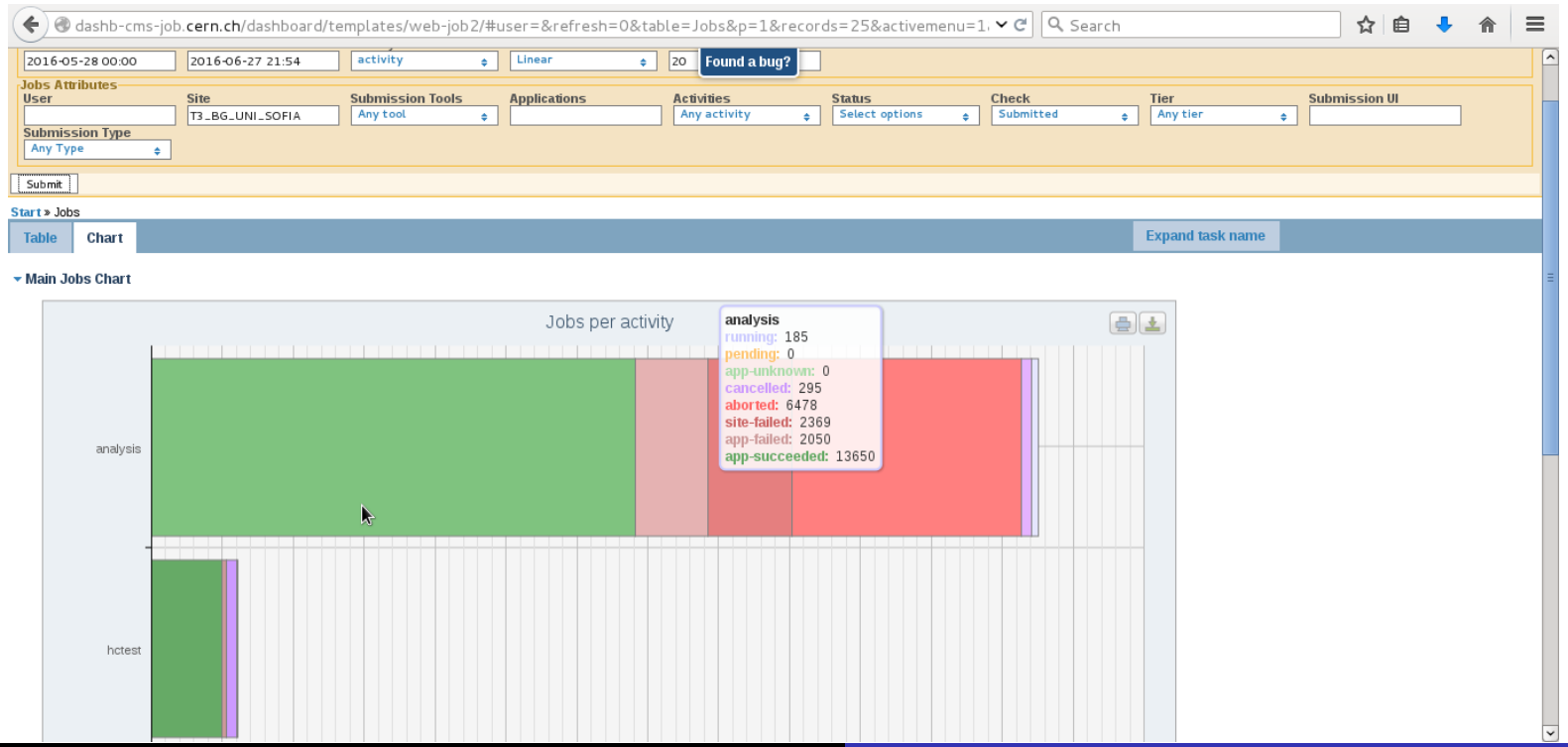
Надеждност на системата



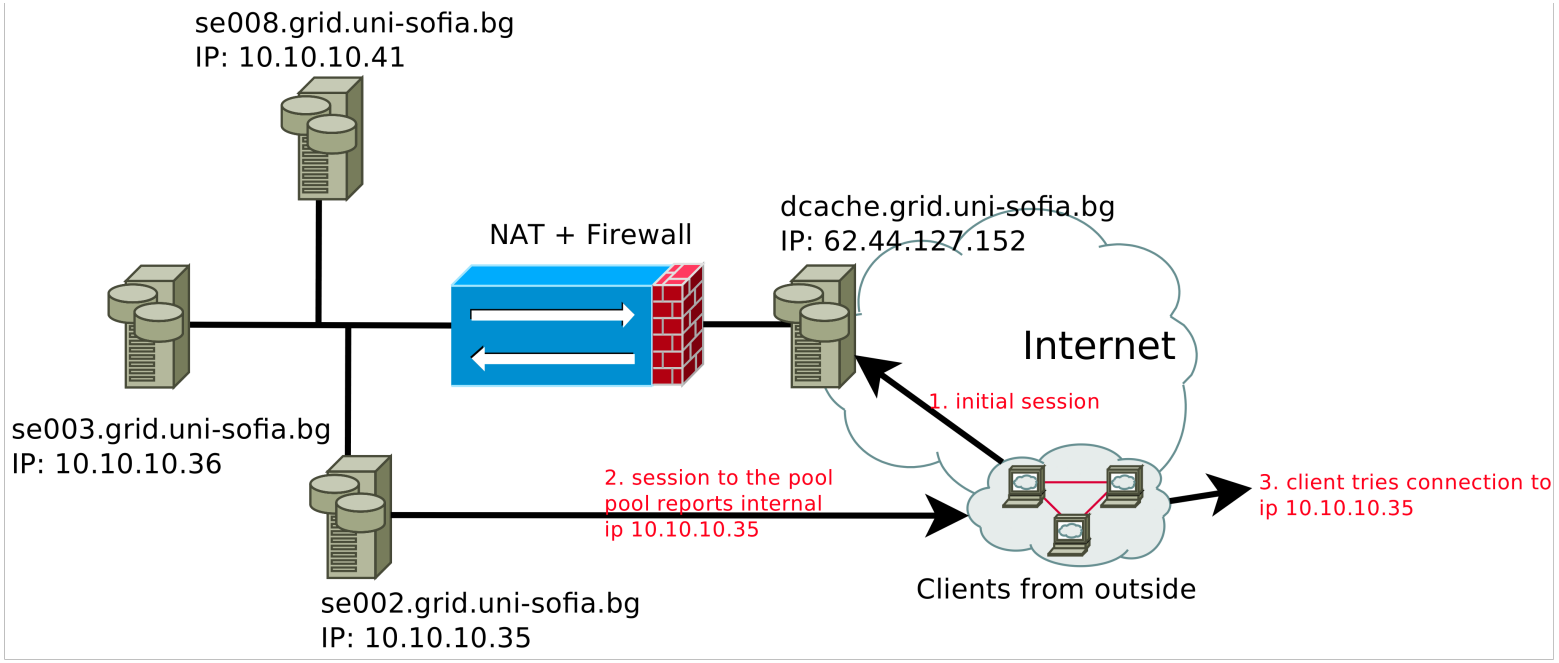
Отработени Изчислителни задачи

The first job record is dated Tue 16 Sep 2014 12:02:47 AM EEST.
The last job record is dated Wed 29 Jun 2016 01:46:20 PM EEST.

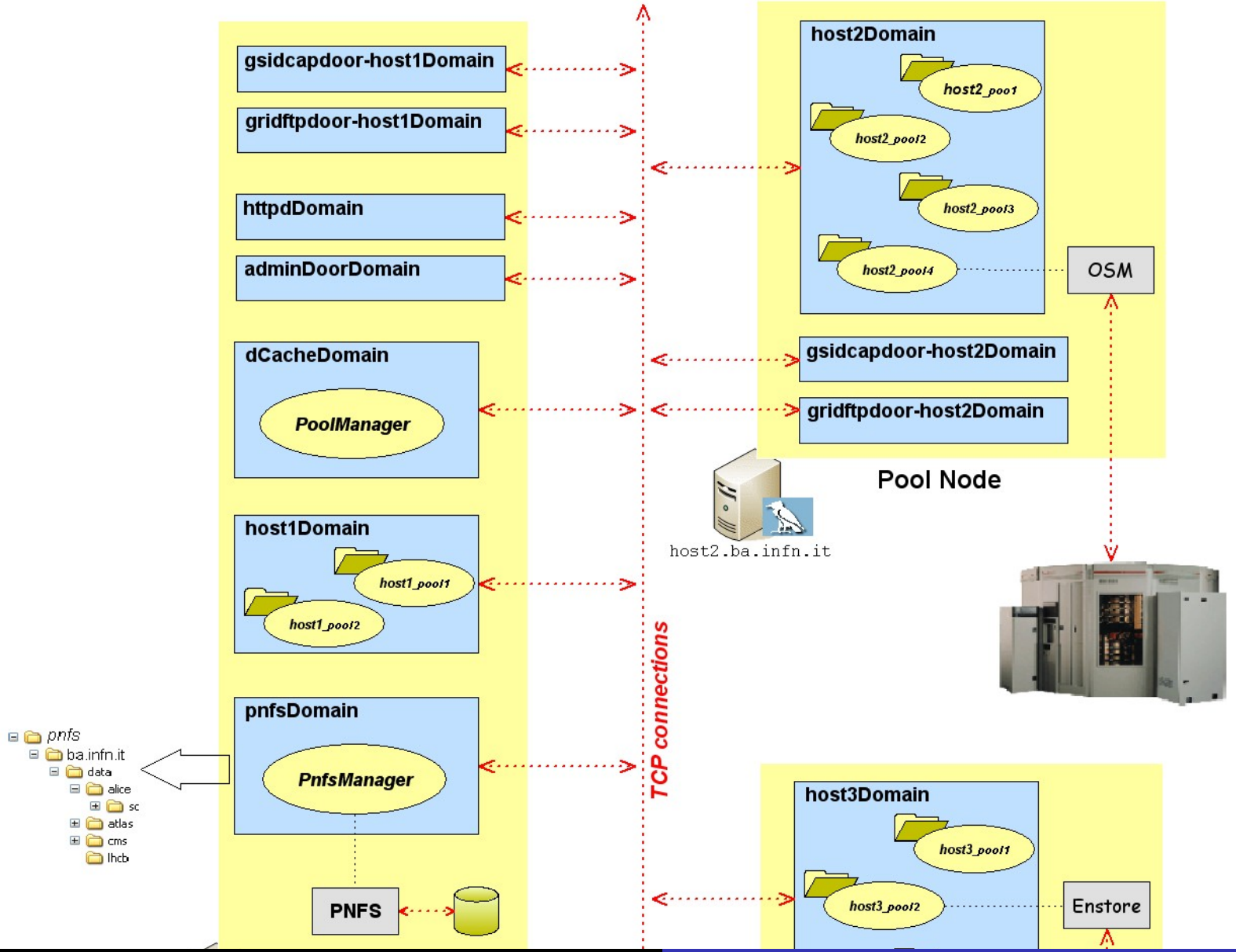
Username	Group	#jobs	days
TOTAL	-	192130	7613.08



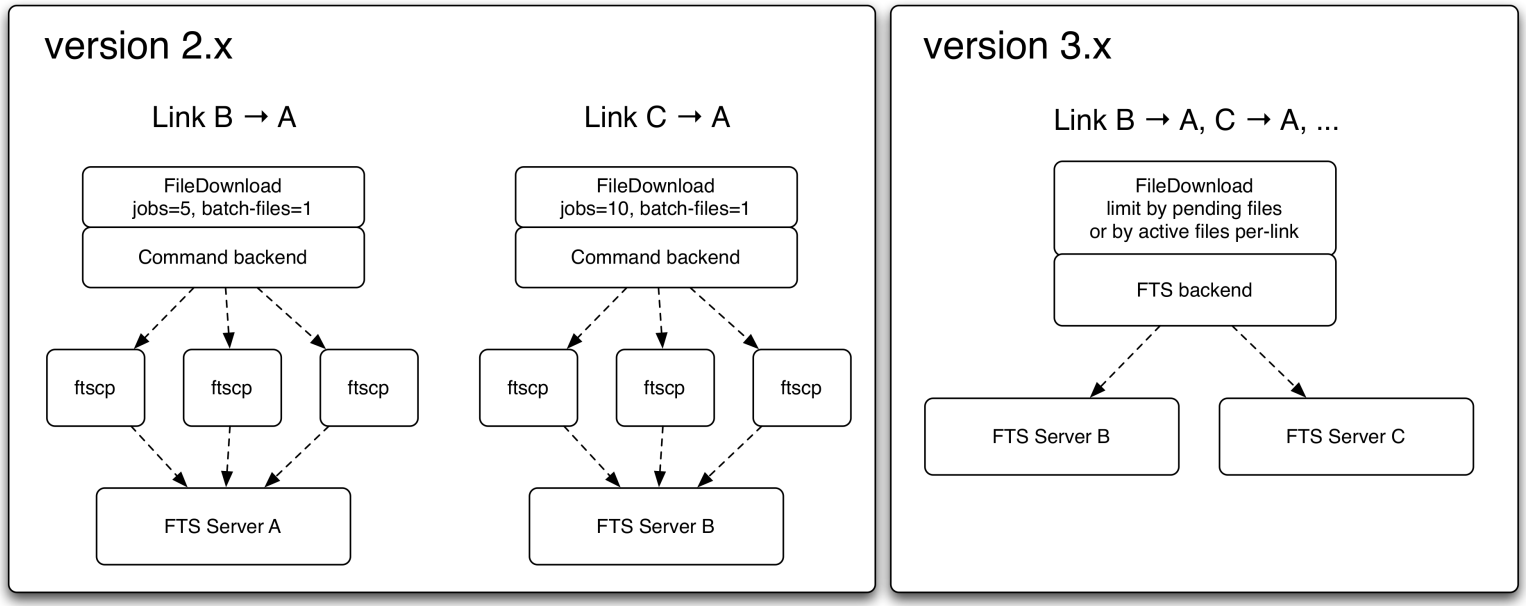
Backup slides



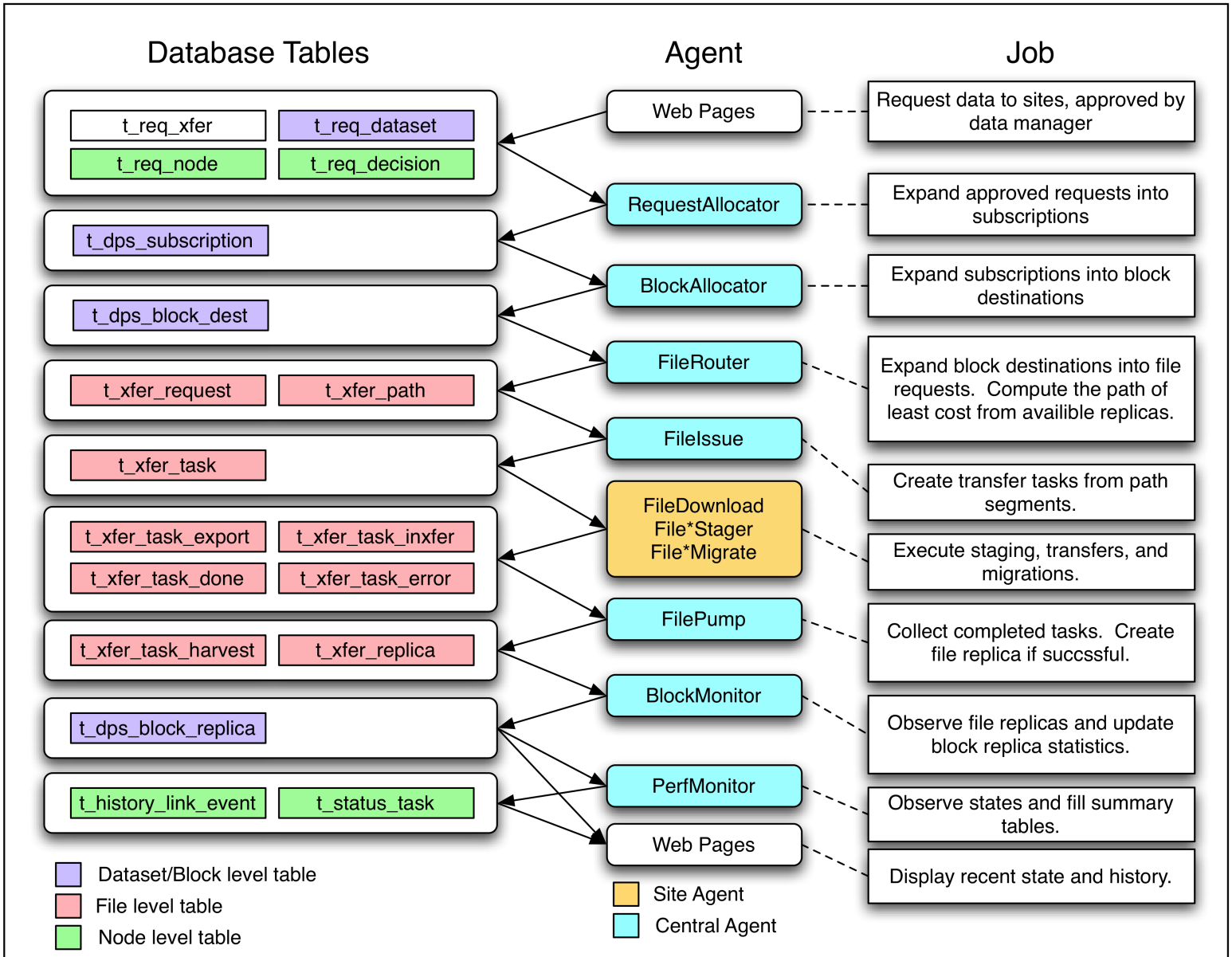
Backup slides



Backup slides



Backup slides



Backup slides

Backup slides

CMS Fill 5339 Report

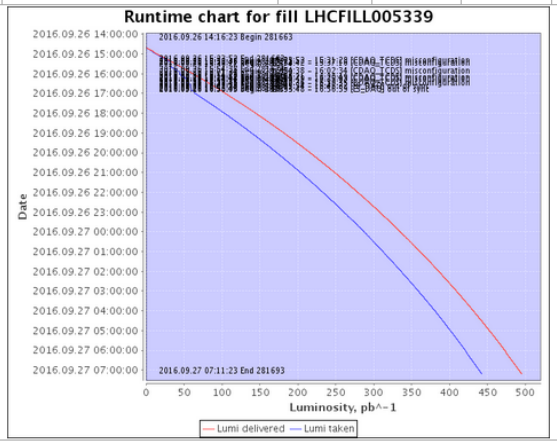
Specific Fill: 5339 GO | Begin: 5339 End: 5339 Stable GO | Last *n* Fills: 20 Stable GO | < 5338 | 5340 > | Show More Options | Help

CMS Fill 5339 Report

BunchFill | LhcEvents | RuntimeLogger | ConditionBrowser | Magnet

CreateTime (declared)	2016.09.26 11:50:46
BeginTime (stable)	2016.09.26 14:40:24
toReady (to HV on)	0.771 *minutes*
toDumpReady	3.617 *minutes*
dumpReadyToDump	1.824 *minutes*
EndTime (dumped)	2016.09.27 07:10:35
Type	*Proton* – PROTON vs PROTON
BField	3.800 T
Energy	6499 GeV
InitialLumi	13062.980×10^{30}cm^{-2}sec^{-1}
PeakLumi	14363.252×10^{30}cm^{-2}sec^{-1}
PeakPileup (interactions/BX)	$\langle n \rangle$ = 46.297
PeakSpecificLumi	5.748×10^{27}cm^{-2}sec^{-1}(10^{11}p)^{2}
DeliveredLumi	495.080 pb^{-1}
RecordedLumi	442.174 pb^{-1}
Efficiency by lumi	89.314%
Efficiency by time	95.302%
Physics Streams Rate	175.804 Hz
InjectionScheme	25ns_2220b_2208_1940_2036_96bpi_24inj
IntensityBeam1	2405.032×10^{11}
IntensityBeam2	2450.361×10^{11}
nBunchesBeam1	2220
nBunchesBeam2	2220
nCollidingBunches	2208
nTargetBunches	2208
CrossingAngle	140.045 µrad
β*	40.0 cm

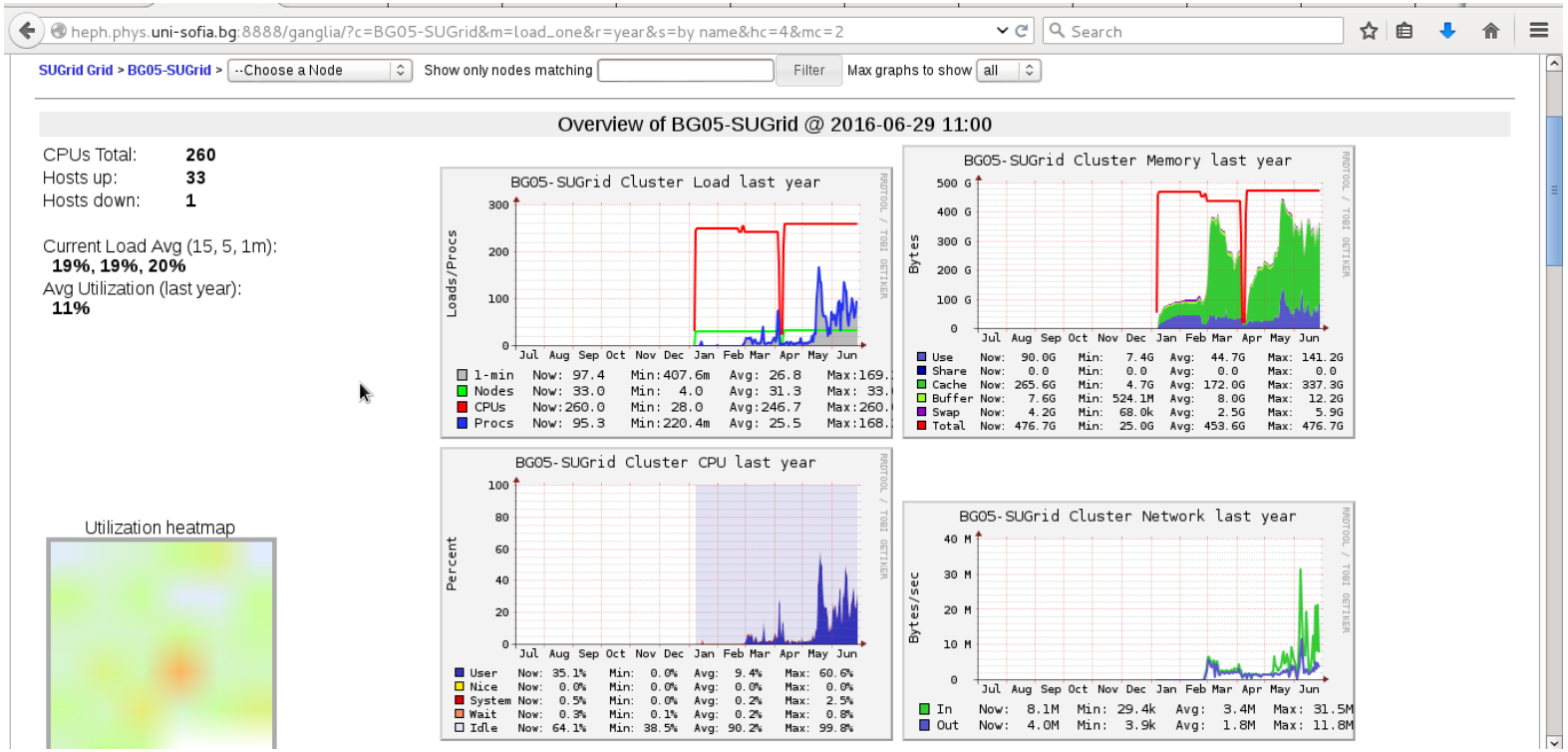
Run	BeginTime	EndTime	Triggers	Lumi pb^{-1}	Recorded pb^{-1}	Eff %
281663	2016.09.26 14:16:23	2016.09.26 15:23:52	214090852	39.677672	31.568352	79.562
281671	2016.09.26 15:30:38	2016.09.26 15:33:41	14433	4.267520	0.002660	0.062
281674	2016.09.26 15:36:22	2016.09.26 15:54:38	70439939	18.872902	12.615572	66.845
281680	2016.09.26 16:01:34	2016.09.26 16:14:20	55880856	13.363340	8.201305	61.372
281685	2016.09.26 16:19:57	2016.09.26 16:21:42	5	2.815156	0.000000	0.000
281686	2016.09.26 16:24:04	2016.09.26 16:32:35	37278349	10.199500	5.245320	51.427
281689	2016.09.26 16:38:46	2016.09.26 16:42:55	10659883	5.969672	1.533633	25.690
281691	2016.09.26 16:47:34	2016.09.26 16:53:48	24327383	5.634297	3.576313	63.474
281693	2016.09.26 16:56:05	2016.09.27 07:11:23	3990958038	392.418148	379.311727	96.660



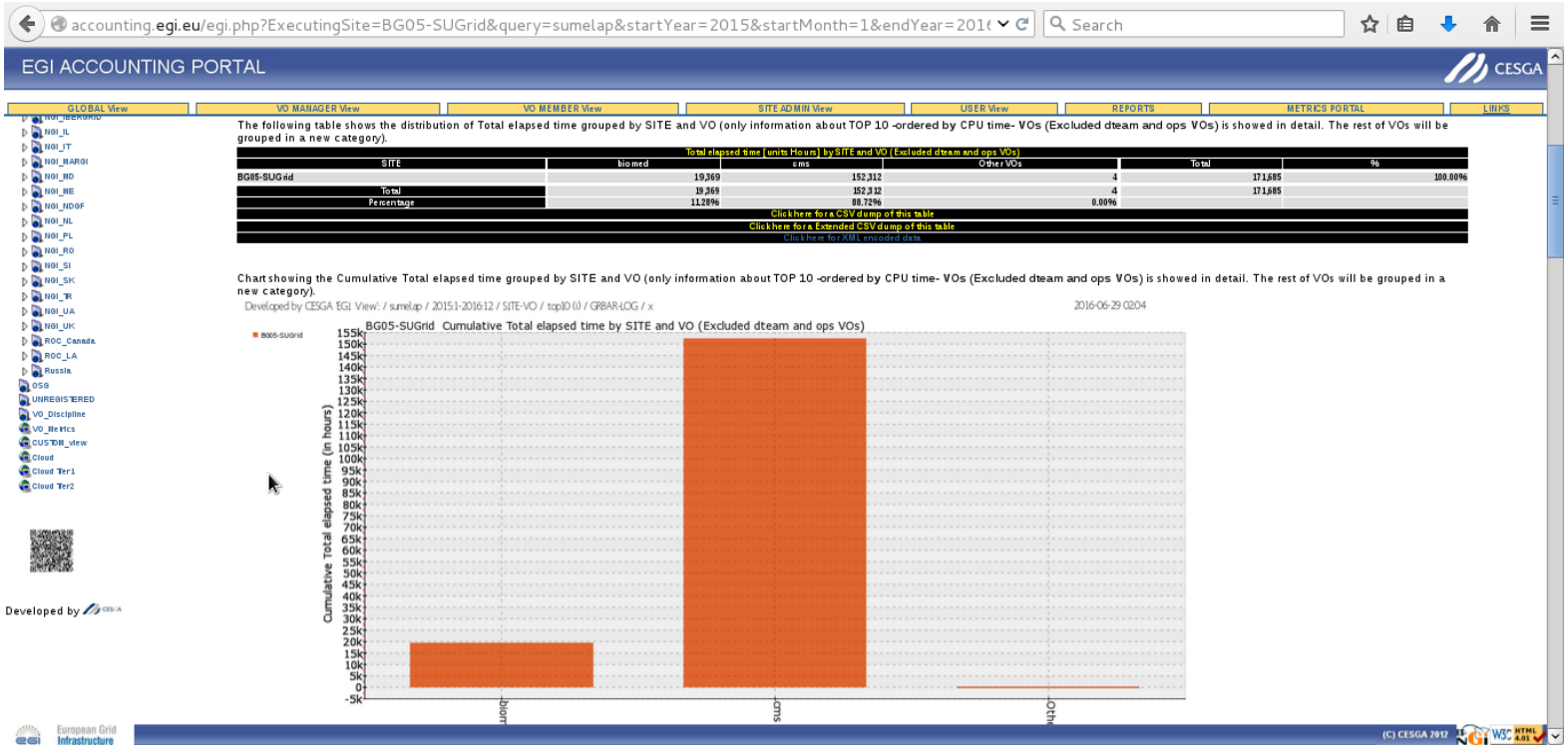
CMS: Fill 5339 Luminosity | CMS: Fill 5339 Instantaneous Luminosity — CMS Online Lumi | CMS: Fill 5339 Lumi per Crossing — Lumi per bx, Spec Lumi per bx

Backup slides

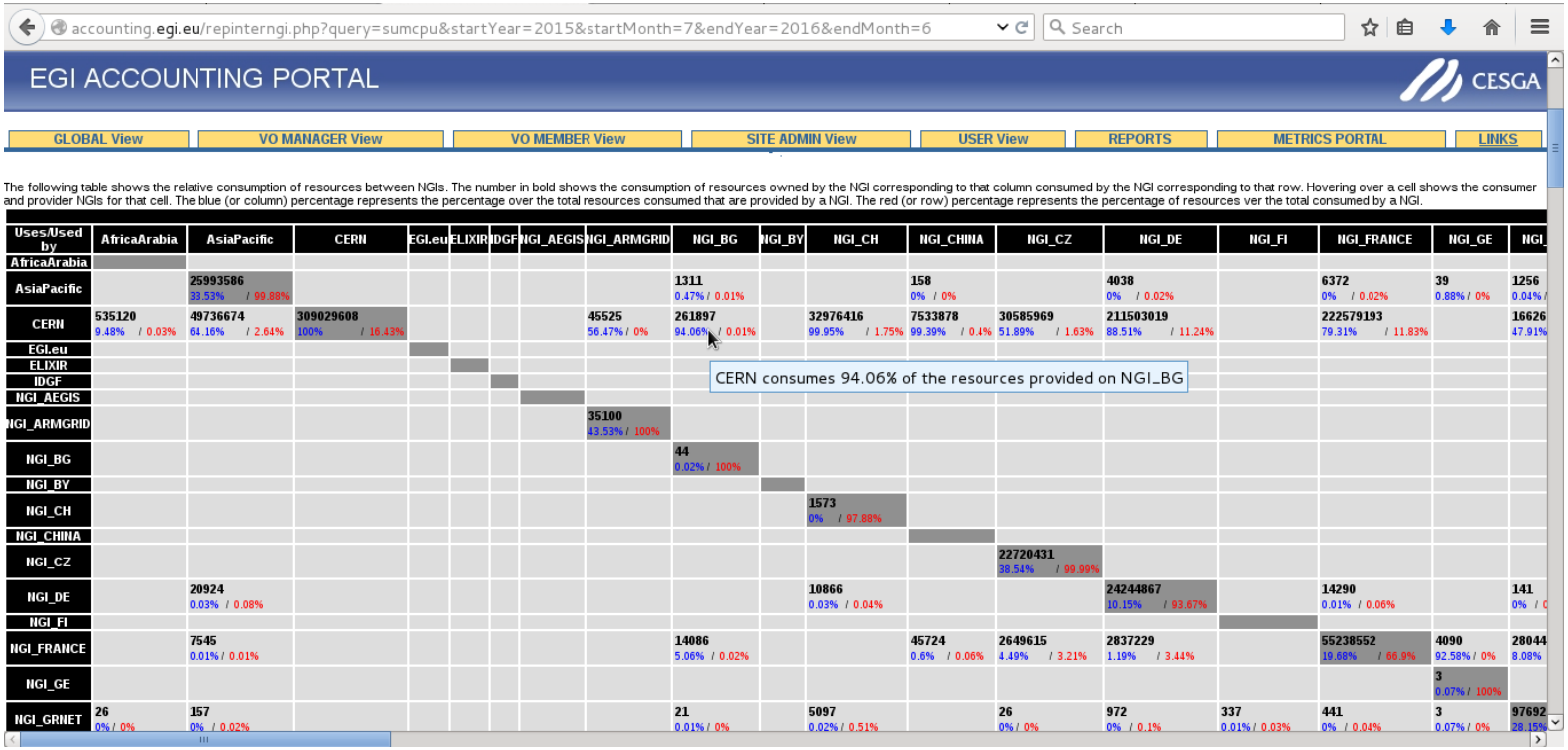
Backup slides



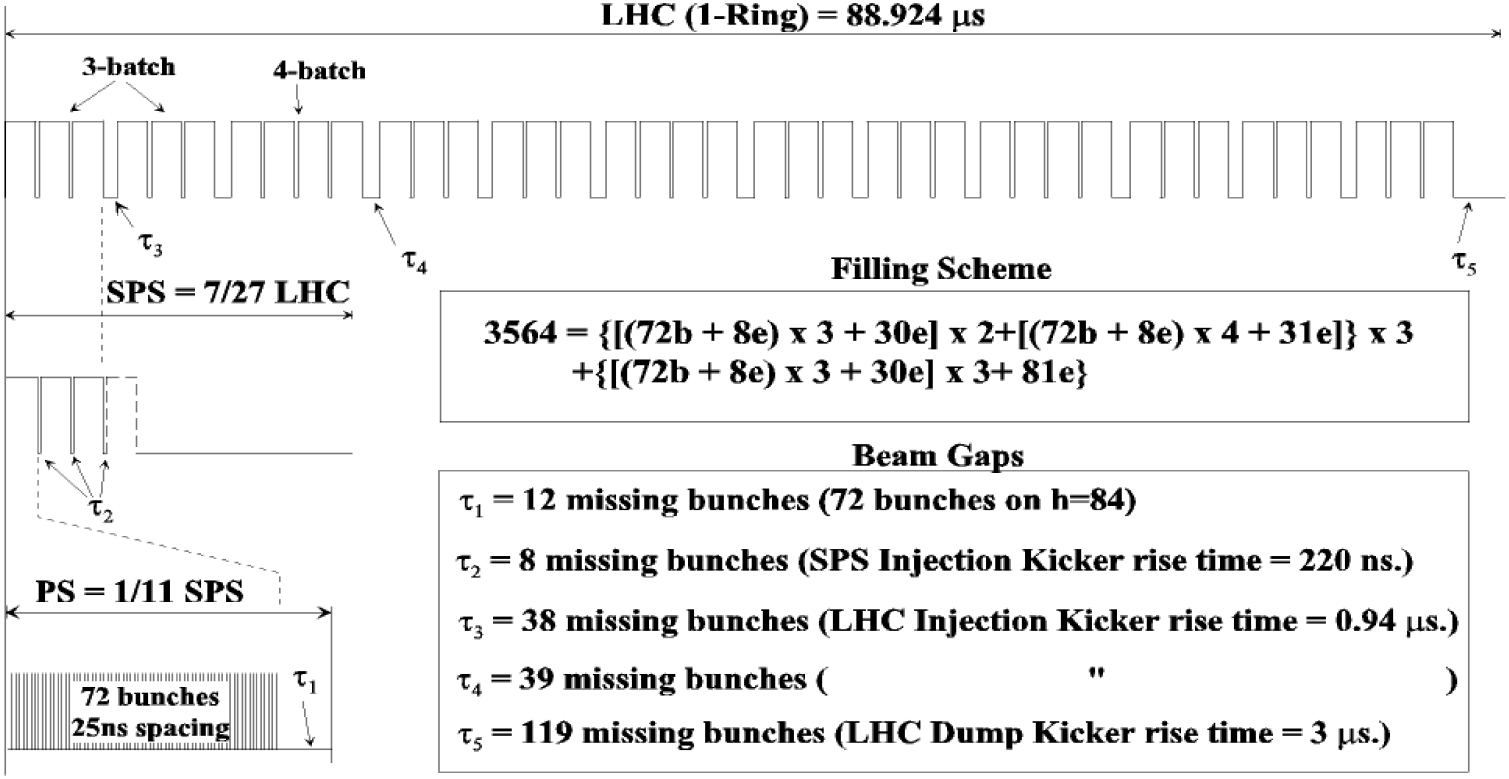
Backup slides



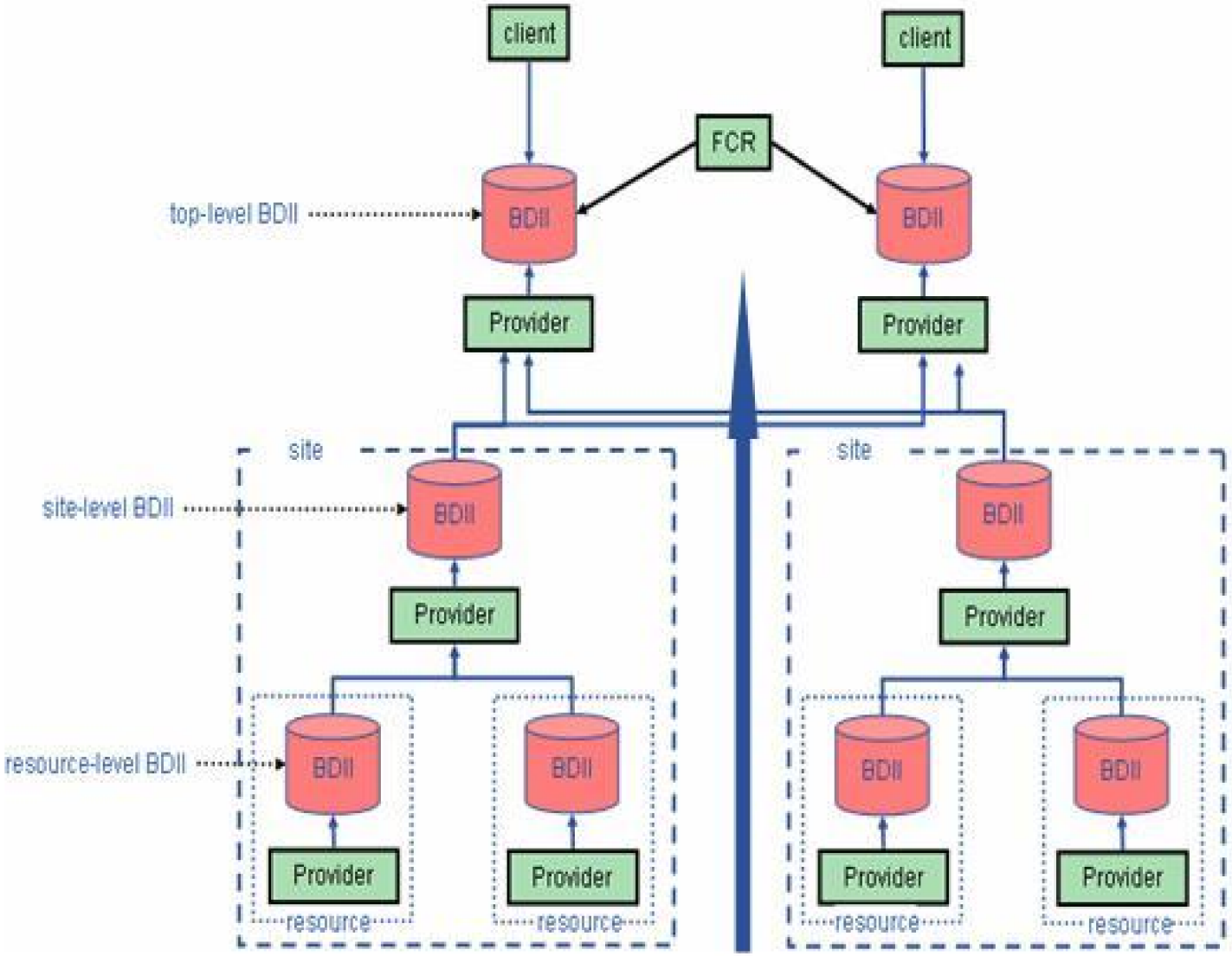
Backup slides



Backup slides



Backup slides



Backup slides