Answers to Panel1/Panel3 Questions

John Harvey/ LHCb

May 12th, 2000

Question 1 : Raw Data Strategy

- Can you be more precise on the term "selected" samples to be exported and the influence of such exported samples (size, who will access, purposes) on networks and other resources.
- Do you plan any back-up ?

Access Strategy for LHCb Data

- Reminder - once we get past the start-up we will be exporting only AOD+TAG data in production cycle, and only ‘small’ RAW+ESD samples (controlled frequency and size) upon physicist request.
- It is a requirement that physicists working remotely should have ~immediate access to AOD+TAG data as soon as they are produced.
  - So as not to penalise physicists working remotely.
  - So as not to encourage physicists working remotely to run their jobs at CERN.
- The rapid distribution of AOD+TAG data from CERN to regional centres places largest load on network infrastructure.

Access to AOD + TAG Data

- EXPORT each ‘week’ to each of the regional centres.
  - AOD+TAG: $5 \cdot 10^7$ events * 20 kb = 1 TB.
- A day turnaround for exporting 1 TB would imply an ‘effective’ network bandwidth requirement of 10 MB/s from CERN to each of the regional centres.
- At the same bandwidth a year’s worth of data could be distributed to the regional centres in 20 days. This would be required following a reprocessing.

Access to RAW and ESD Data

- Worst case planning- the start-up period (1st and 2nd years)
  - We export sufficient RAW + ESD data to allow for remote physics and detector studies
- At start-up our selection tagging will be crude, so assume we select 10% of data taken for a particular channel, and of this sample include RAW+ESD for 10% of events
- EXPORT each 'week' to each of the regional centres
  - AOD+TAG: 5 . 10^7events * 20 kb = 1 TB
  - RAW+ESD: 10 channels * 5.10^5 events * 200 KB =1TB
- A day turnaround for exporting 2 TB implies 'effective' bandwidth requirement of 20 MB/s from CERN to each of RCs

Access to RAW and ESD Data

- In steady running after the first 2 years people will still want to access the RAW/ESD data for physics studies but only for their private selected sample.
- The size of the data samples required in this case is small,
  - of the order of 10^5 events (i.e. 20 GB) per sample
  - turnaround time < 12 hours
  - bandwidth requirement for one such transaction is <1MB/s

Access to RAW and ESD Data

- Samples of RAW and ESD data will also be used to satisfy requirements for detailed detector studies.
- Samples of background events may also be required, but it is expected that the bulk of the data requirements can be satisfied with the samples distributed for physics studies.
- It should be noted however that for detector/trigger studies people working on detectors will most likely be at CERN, and it may not be necessary to export RAW+ESD for such studies during the start-up period.
- After first two years smaller samples will be needed for detailed studies of detector performance.

Backup of Data

- We intend to make to two copies of RAW data on archive media (tape)

Question 2 : Simulation

- Can you be more precise about your MDC (mock data challenges) strategy ?
- In correlation with hardware cost decreases. (Remember : a 10% MDC 3 years before T° could cost ~ as much as a 100% MDC at T°)

Physics : Plans for Simulation 2000-2005

- In 2000 and 2001 we will produce 3. 10^6 simulated events each year for detector optimisation studies in preparation of the detector TDRs (expected in 2001 and early 2002).
- In 2002 and 2003 studies will be made of the high level trigger algorithms for which we are required to produce 6.10^6 simulated events each year.
- In 2004 and 2005 we will start to produce very large samples of simulated events, in particular background, for which samples of 10^7 events are required.
- This on-going physics production work will be used as far as is practicable for testing development of the computing infrastructure.

Computing : MDC Tests of Infrastructure

- 2002 : MDC 1 - application tests of grid middleware and farm management software using a real simulation and analysis of 10^7 B channel decay events. Several regional facilities will participate :
  - CERN, RAL, Lyon/CCIN$_2$P$_3$,Liverpool, INFN, ...
- 2003 : MDC 2 - participate in the exploitation of the large scale Tier0 prototype to be setup at CERN
  - High Level Triggering – online environment, performance
  - Management of systems and applications
  - Reconstruction – design and performance optimisation
  - Analysis – study chaotic data access patterns
  - STRESS TESTS of data models, algorithms and technology
- 2004 : MDC 3 - Start to install event filter farm at the experiment to be ready for commissioning of detectors in 200 4 and 2005

Growth in Requirements to Meet Simulation Needs

	2000	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*
No of signal events	1.0E+06	1.0E+06	2.0E+06	3.0E+06	5.0E+06	1.0E+07	1.0E+07	1.0E+07	1.0E+07	1.0E+07	1.0E+07
No of background events	1.0E+06	1.5E+06	2.0E+06	4.0E+06	1.0E+07	1.0E+09	1.0E+09	1.0E+09	1.0E+09	1.0E+09	1.0E+09
CPU for simulation of signal (SI9	10000	10000	20000	30000	50000	100000	100000	100000	100000	100000	100000
CPU for background simulation (	16000	24000	32000	64000	160000	400000	400000	400000	400000	400000	400000
CPU user analysis (SI95)	2500	2500	5000	7500	12500	25000	25000	25000	25000	25000	25000
RAWmc data on disk (TB)	0.4	0.5	0.8	1.4	3	202	202	202	202	202	202
RAWmc data on tape (TB)	0.4	0.5	0.8	1.4	3	202	202	202	202	202	202
ESDmc data on disk (TB)	0.2	0.25	0.4	0.7	1.5	101	101	101	101	101	101
AODmc data on disk (TB)	0.06	0.1	0.1	0.3	0.5	30.5	39.4	42.1	42.9	43.2	43.3
TAGmc data on disk (TB)	0.002	0.0025	0.004	0.007	0.015	1.01	1.01	1.01	1.01	1.01	1.01

Unit Costs

	2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
CPU cost / SI95	64.5	46.1	32.9	23.5	16.8	12.0	8.6	6.1	4.4	3.1	2.2
Disk cost / GB	16.1	11.5	8.2	5.9	4.2	3.0	2.1	1.5	1.1	0.8	0.6
Tape cost / GB	2.7	1.9	1.4	1.0	0.7	0.5	0.36	0.26	0.18	0.13	0.09

	2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
CPU for signal(kSFr)	323	230	329	235	336	600	171	122	87	62	45
CPU for background (kSFr)	0	369	263	753	1613	2880	857	612	437	312	223
CPU for user analysis (kSFr)	65	92	66	59	84	150	69	49	35	25	18
RAWmc data on disk (kSFr)	3	6	2	4	7	597	129	92	66	47	33
RAWmc data on tape (kSFr)	0.2	0.2	0.2	0.3	0.4	20.2	14.4	10.3	7.4	5.3	3.8
ESDmc data on disk (kSFr)	3	3	1	2	3	299	64	46	33	23	17
AODmc data on disk (kSFr)	1	1	0	1	1	90	21	15	11	8	6
TAGmc data on disk (kSFr)	0.0	0.0	0.0	0.0	0.1	3.0	2.2	1.5	1.1	0.8	0.6
Investment per year (kSFr)	**395**	**701**	**663**	**1053**	**2045**	**4639**	**1328**	**949**	**678**	**484**	**346**

Cost / Regional Centre for Simulation

- Assume there are 5 regional centres
- Assume costs are shared equally

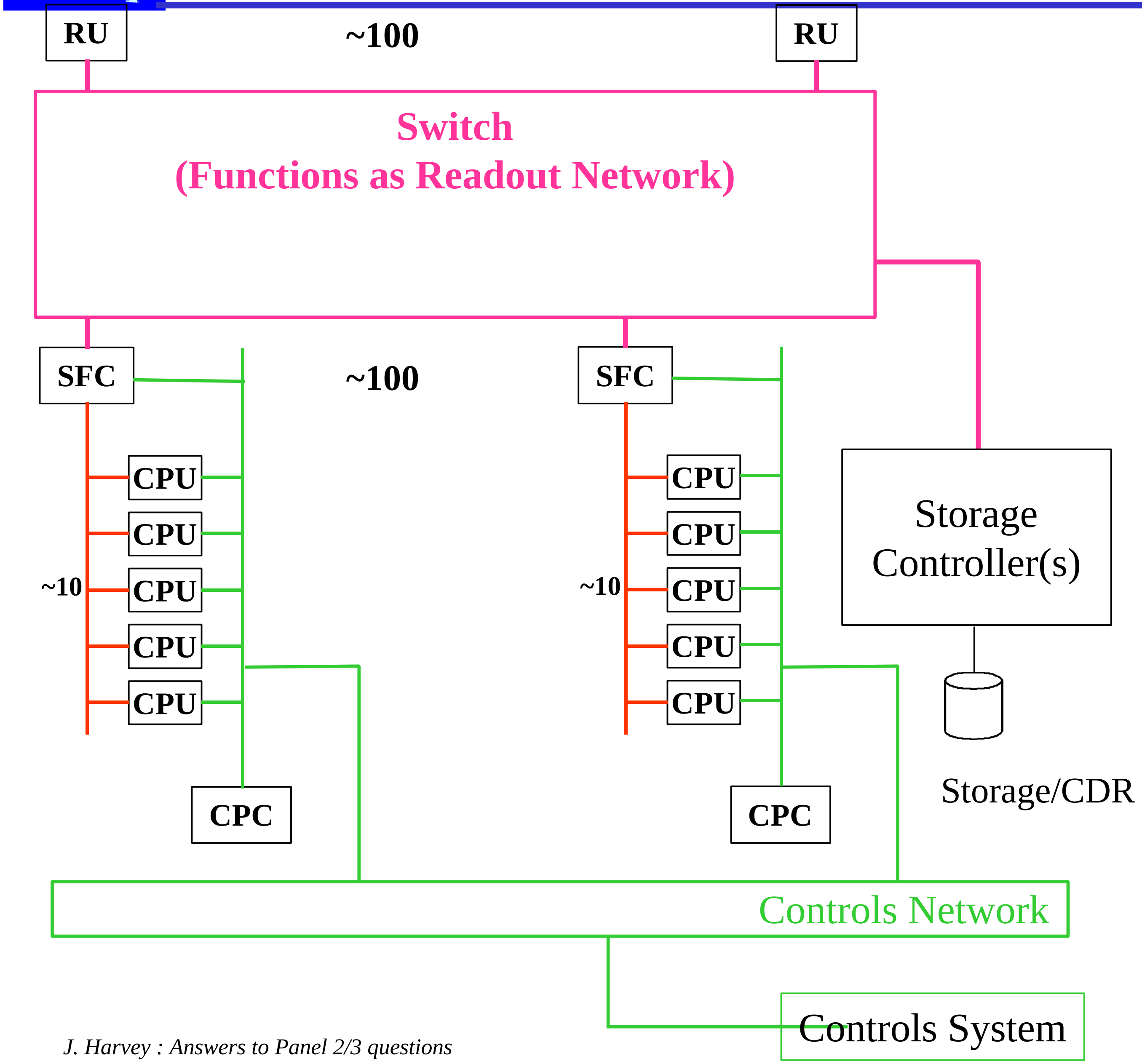
	2000	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*
No of signal events	2.0E+05	2.0E+05	4.0E+05	6.0E+05	1.0E+06	2.0E+06	2.0E+06	2.0E+06	2.0E+06	2.0E+06	2.0E+06
No of background events	2.0E+05	3.0E+05	4.0E+05	8.0E+05	2.0E+06	2.0E+08	2.0E+08	2.0E+08	2.0E+08	2.0E+08	2.0E+08
CPU for simulation of signal (SI9	2000	2000	4000	6000	10000	20000	20000	20000	20000	20000	20000
CPU for simulation of backgroun	3200	4800	6400	12800	32000	80000	80000	80000	80000	80000	80000
CPU user analysis (SI95)	500	500	1000	1500	2500	5000	5000	5000	5000	5000	5000
RAWmc data on disk (TB)	0.08	0.1	0.16	0.28	0.6	40.4	40.4	40.4	40.4	40.4	40.4
RAWmc data on tape (TB)	0.08	0.1	0.16	0.28	0.6	40.4	40.4	40.4	40.4	40.4	40.4
ESDmc data on disk (TB)	0.04	0.05	0.08	0.14	0.3	20.2	20.2	20.2	20.2	20.2	20.2
AODmc data on disk (TB)	0.012	0.0186	0.02958	0.05087	0.10526	6.09158	7.88747	8.42624	8.58787	8.63636	8.65091
TAGmc data on disk (TB)	0.0004	0.0005	0.0008	0.0014	0.003	0.202	0.202	0.202	0.202	0.202	0.202

	2000	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
CPU for signal (kSFr)	64.5	46.1	65.9	47.0	67.2	120.0	34.3	24.5	17.5	12.5	8.9
CPU for background (kSFr)	0.0	73.8	52.7	150.5	322.6	576.0	171.4	122.4	87.5	62.5	44.6
CPU user analysis (kSFr)	12.9	18.4	13.2	11.8	16.8	30.0	13.7	9.8	7.0	5.0	3.6
RAWmc data on disk (kSFr)	0.6	1.2	0.5	0.7	1.3	119.4	25.7	18.4	13.1	9.4	6.7
RAWmc data on tape (kSFr)	0.0	0.0	0.0	0.1	0.1	4.0	2.9	2.1	1.5	1.1	0.8
ESDmc data on disk (kSFr)	0.6	0.6	0.2	0.4	0.7	59.7	12.9	9.2	6.6	4.7	3.3
AODmc data on disk (kSFr)	0.2	0.2	0.1	0.1	0.2	18.0	4.3	3.1	2.2	1.6	1.1
TAGmc data on disk (kSFr)	0.0	0.0	0.0	0.0	0.0	0.6	0.4	0.3	0.2	0.2	0.1
Investment per year (kSFr)	**79**	**140**	**133**	**211**	**409**	**928**	**266**	**190**	**136**	**97**	**69**

Tests Using Tier 0 Prototype in 2003

- We intend to make use of the Tier 0 prototype planned for construction in 2003 to make stress tests of both hardware and software
- We will prepare realistic examples of two types of application :
  - Tests designed to gain experience with the online farm environment
  - Production tests of simulation, reconstruction, and analysis

Event Filter Farm Architecture



RU ~100 RU

Switch
(Functions as Readout Network)

SFC ~100 SFC

CPU
CPU
~10 CPU
CPU
CPU

CPC

Storage
Controller(s)

Storage/CDR

Controls Network

Controls System

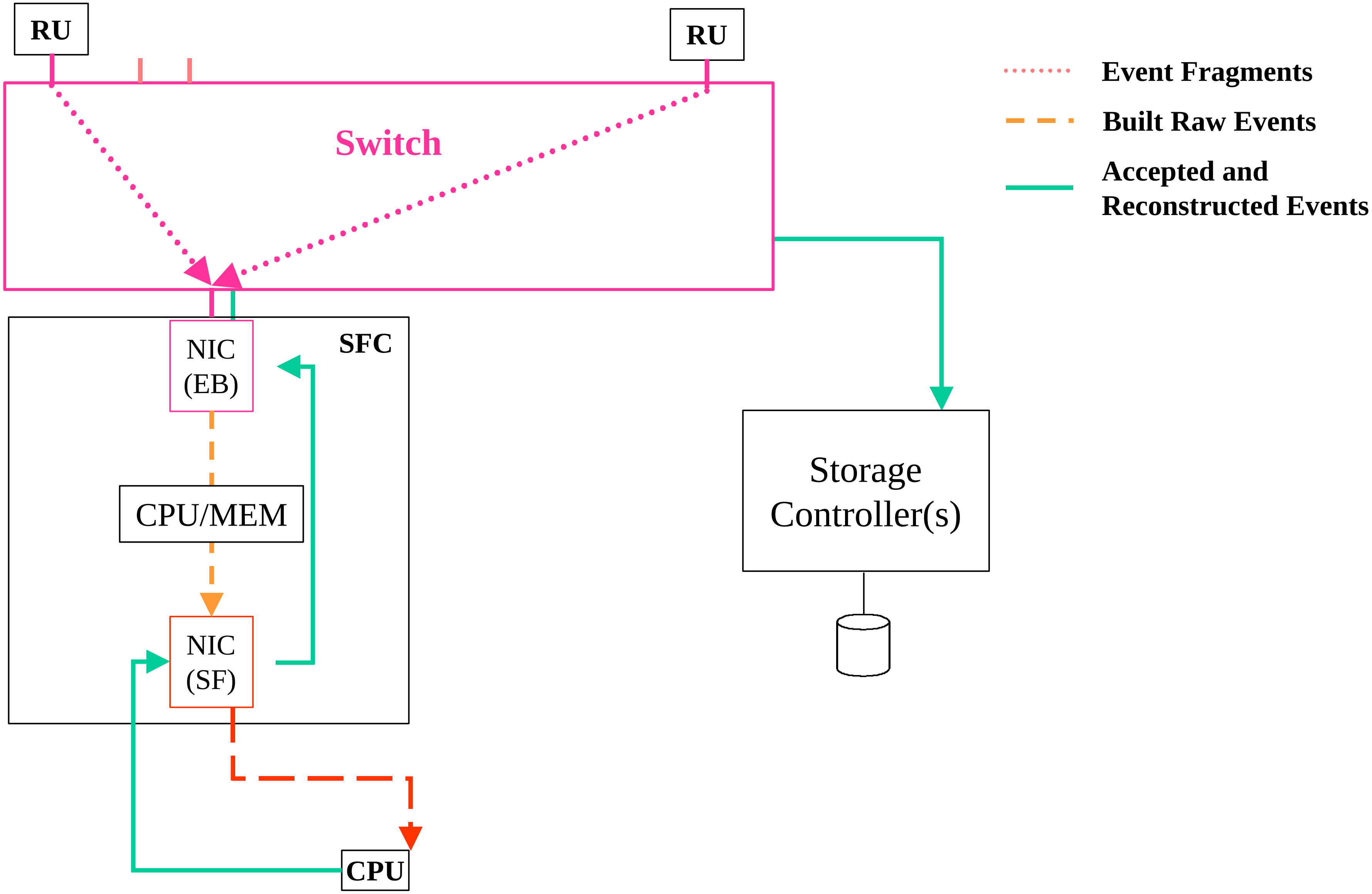
Readout Network
Technology (GbE?)

Sub-Farm Network
Technology
(Ethernet)

Controls Network
Technology
(Ethernet)

SFC	**Sub-Farm Controller**
CPC	**Control PC**
CPU	**Work CPU**

Data Flow



RU

RU

Switch

SFC

NIC (EB)

CPU/MEM

NIC (SF)

CPU

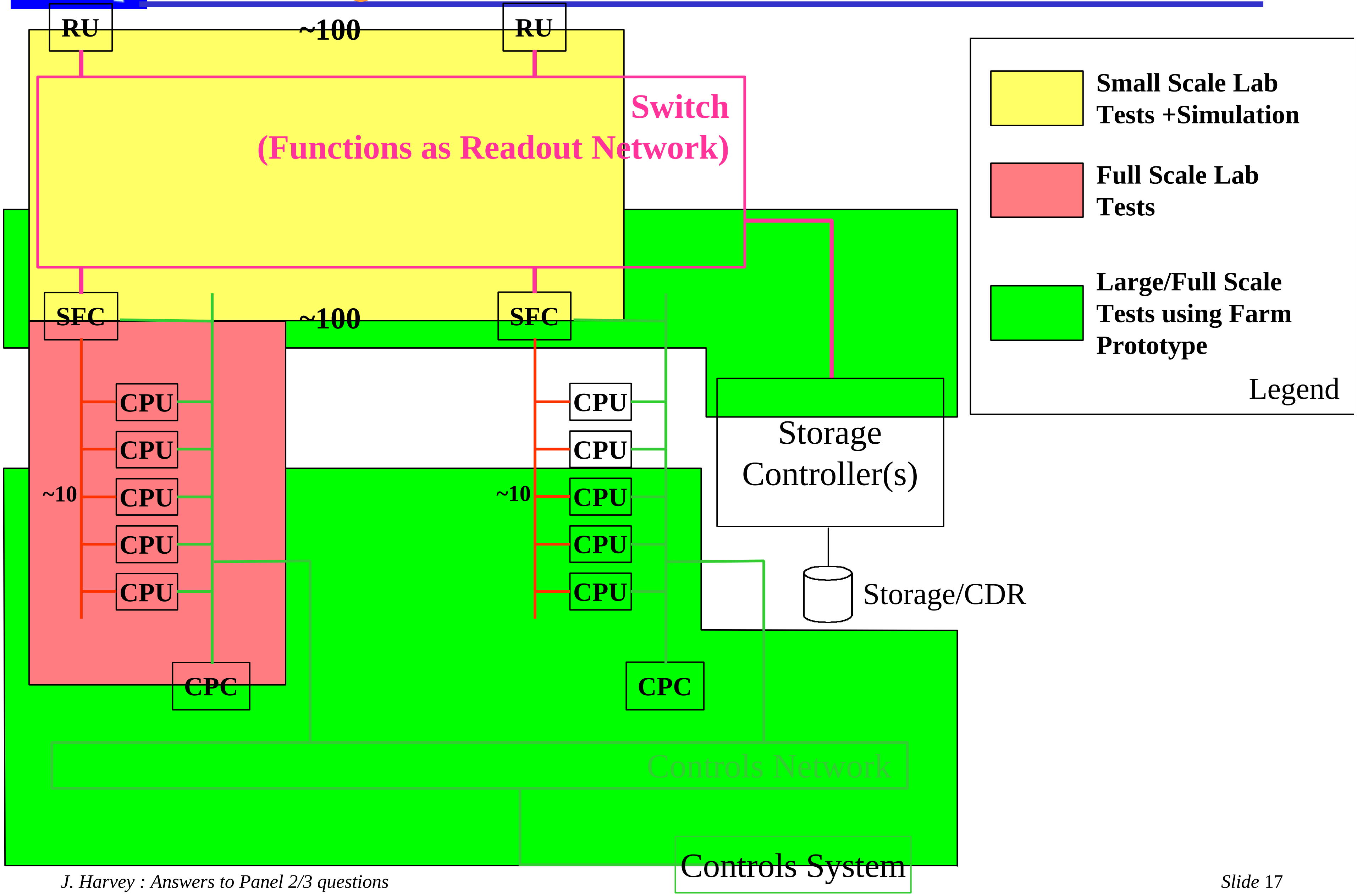
Storage Controller(s)

Event Fragments

Built Raw Events

Accepted and Reconstructed Events

Testing/Verification



RU

~100

RU

Switch

(Functions as Readout Network)

SFC

~100

SFC

CPU

CPU

~10

CPU

CPU

CPU

CPC

CPU

CPU

~10

CPU

CPU

CPU

CPC

Storage Controller(s)

Storage/CDR

Controls Network

Controls System

🟨 Small Scale Lab Tests +Simulation

🟥 Full Scale Lab Tests

🟩 Large/Full Scale Tests using Farm Prototype

Legend

Requirements on Farm Prototype

- Functional requirements
  - A separate controls network (Fast Ethernet at the level of the sub-farm, GbEthernet towards the controls system)
  - Farm CPUs organized in sub-farms (contrary to a "flat" farm)
  - Every CPU in the sub-farm should have two Fast Ethernet interfaces
- Performance and Configuration Requirements
  - SFC : NIC >1 MB/s, >512 MB Memory
  - Storage controller : NIC >40-60 MB/s, >2 GB memory, >1 TB disk
  - Farm CPU : ~256 MB memory
  - Switch : >95 ports @ 1 Gb/s (Gbit Ethernet)

Data Recording Tests

- Raw and reconstructed data are sent from ~100 SFCs to the storage controller and inserted in the permanent storage in a format suitable for re-processing and off-line analysis.
- Performance Goal
  - The storage controller should be able to populate the permanent storage at a event rate of ~200 HZ and an aggregate data rate of ~40-50 MB/s
- Issues to be studied
  - Data movement compatible with DAQ environment
  - Scalability of Data Storage

Farm Controls Tests

- A large farm of processors are to be controlled through a controls system
- Performance Goal
  - Reboot all farm CPUs in less than ~10 minutes
  - configure all Farm CPUs in less than ~1 minute
- Issues to be studied
  - Scalability of booting method
  - Scalability of controls system
  - Scalability of access and distribution of configuration data

Scalability tests for simulation and reconstruction

- Test writing of reconstructed+raw data at 200Hz in online farm environment
- Test writing of reconstructed+simulated data in offline Monte Carlo farm environment
  - Population of event database from multiple input processes
- Test efficiency of event and detector data models
  - Access to conditions data from multiple reconstruction jobs
  - Online calibration strategies and distribution of results to multiple reconstruction jobs
  - Stress testing of reconstruction to identify hot spots, weak code etc.

Scalability tests for analysis

- Stress test of event database
  - Multiple concurrent accesses by “chaotic” analysis jobs

- Optimisation of data model
  - Study data access patterns of multiple, independent, concurrent analysis jobs
  - Modify event and conditions data models as necessary
  - Determine data clustering strategies

Question 3 : Luminosity and Detector Calibration

- Strategy in the analysis to get access to the conditions data.
- Will it be performed at CERN only or at outside institutes.
- If outside,how the raw data required can be accessed and how the detector conditions DB will be updated?

Access to Conditions Data

- Production updating of conditions database (detector calibration) to be done at CERN for reasons of system integrity.
- Conditions data less than 1% of event data
- Conditions data for relevant period will be exported as part of the production cycle to the Regional Centres .
  - Detector status data being designed
    - < 100 kbyte/sec ~ < 10 GB/week
  - Essential Alignment + calibration constants required for reconstruction
    - ~ 100 MB/week

Luminosity and Detector Calibration

- Comments on detector calibration
  - VELO done online ..needed for trigger(pedestals,common mode + alignment for each fill)
  - Tracking alignment will be partially done at start-up without magnetic field
  - CALORIMETER done with test beam and early physics data
  - RICHs will have optical alignment system

- Comment on luminosity calibration(based at CERN)
  - Strategy being worked on. Thinking to base on 'number of primary vertices' distribution (measured in an unbiased way)

Question 4 : CPU estimates

- "floating" factors, at least 2, were quoted at various meetings by most experiments. And the derivative is definitely positive. Will your CPU estimates continue to grow ?
- How far ?
- Are you convinced your estimates are right within a factor 2 ?
- Would you agree with a CPU sharing of 1/3, 1/3, 1/3 between Tier0,{Tier1},{Tier2,3,4} ?

CPU Estimates

- CPU estimates have been made using performance measurements made with today's software
- Algorithms have still to be developed and final technology choices made e.g.for data storage, …
- Performance optimisation will help reduce requirements
- Estimates will be continuously revised
- The profile with time for acquiring cpu and storage has been made.
- Following acquisition of the basic hardware it assume that acquisition will proceed at 30% each year for cpu and 20% for disk. This is to cover growth and replacement.
- We will be limited by what is affordable and will adapt our simulation strategy accordingly

Question 5 : Higher network bandwidth

- Please summarise the bandwidth requirements associated with the different elements of the current baseline model. Also please comment on possible changes to the model if very high, guaranteed bandwidth links (10 Gbps) become available.

NB. With a 10Gbps sustained throughput (ie. a ~20G link), one could transfer

- a 40 GB tape in half a minute,
- one TB in less than 15',
- one PB in 10 days.

Bandwidth requirements in/out of CERN

	2004	***2005***	***2006***	***2007***	***2008***	***2009***	***2010***
Export of AOD and TAG real data (MB/s)		50	50	50	50	50	50
Export of RAW and ESD real data (MB/s)		50	50	1	1	1	1
Export of conditions data (MB/s)		0.1	0.1	0.1	0.1	0.1	0.1
Import of AOD and TAG simulated data (MB/s)	50	50	50	50	50	50	50
Total bandwidth (MB/s)	**50**	**150**	**150**	**101**	**101**	**101**	**101**

Impact of 10 Gbps connections

- The impact of very high bandwidth network connections would be to give optimal turnround for the distribution of AOD and TAG data and to give very rapid access to RAW and ESD data for specific physics studies.
- Minimising the latency for response to individual physicist requests is convenient and improves efficiency of analysis work
- At present we do not see any strong need to distribute all the RAW and ESD data as part of the production cycle
- We do not rely on this connectivity but will exploit it if it is affordable.

Question 6 : Event storage DB management tools

- Options include Objectivity, Root, a new project Espresso or an improved version of the first two ?
- Shall we let experiments make a free choice or be more directive ?
- Shall we encourage a commercial product or an in-house open software approach ?
- Multiple choice would mean less resources per experiment. Can we afford to have different such tools for the 4 experiments ? only one, two maximum, can we interfere with decisions in which each experiment has already invested many man-years or shall we listen more to the "all purpose Tier1" (= a Tier-1 that will support several LHC experiments, plus perhaps non LHC experiments) that would definitely prefer a support to a minimum of systems? Similar comments could be made about other software packages.

Free choice?

- The problem of choice is not only for the DB management tool. The complete data handling problem needs to be studied and decisions need to be made.
- This comprises the object persistency and its connection to the experiment framework, bookkeeping and event catalogs, interaction with the networks and mass storage, etc.
- It involves many components machines, disks, robots, tapes, networks, etc. and a number of abstraction layers.
- The choice of the product or solution for each of the layers needs to be carefully studied as a coherent solution.

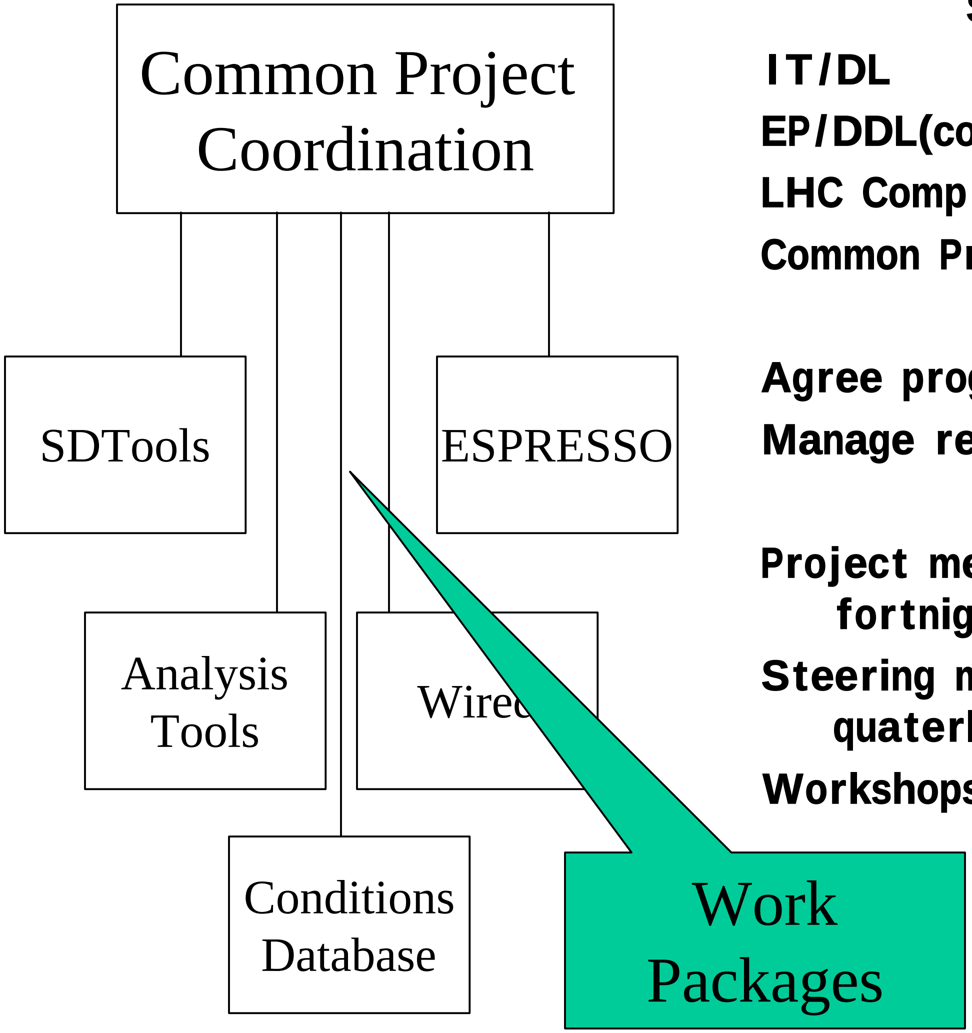
Commercial or in-house

- We are of the opinion that more than one object storage solution should be available to the LHC experiments. Each one with a different range of applicability.
  - a full-fledged solution for the experiment main data store capable of storing petabytes distributed worldwide – implies security, transactions, replication, etc. (commercial)
  - a much lighter solution for end-physicists doing the final analysis with his own private dataset. (in-house)
- Perhaps a single solution can cover the complete spectrum but in general this would not the case.
- If commercial solution is not viable then an in-house solution will have to be developed

Question 7 : Coordination Body?

- A complex adventure such as the LHC computing needs a continuous coordination and follow-up at least until after the first years of LHC running.
- What is your feeling on how this coordination should be
- organized ?
- How would you see a "LCB++" for the coming decade ?

LCB++ - a possible scenario



Steering

IT/DL

EP/DDL(computing)

LHC Comp. Coordinators

Common Project Coordinator

Review

Independent reviewers

Report to management

(Directors, spokesmen,..)

Agree programme

Manage resources

Project meetings / fortnightly

Steering meetings / quaterly

Workshops / quaterly

Follows structure of JCOP