Baseline Model of LHCb's Distributed Computing Facilities

Report to World Wide Analysis Review Panel

23 March 2000

J.Harvey / CERN

Talk Outline

- Logical dataflow and workflow model
- Data processing and storage requirements
- Baseline computing model
- Comparison with MONARC generic model
- Data processing at CERN
- Plans for deployment of the computing model

Following talk by Frank Harris will focus on :

- Resources at regional facilities
- Comments on Disk vs Tape

General Comments

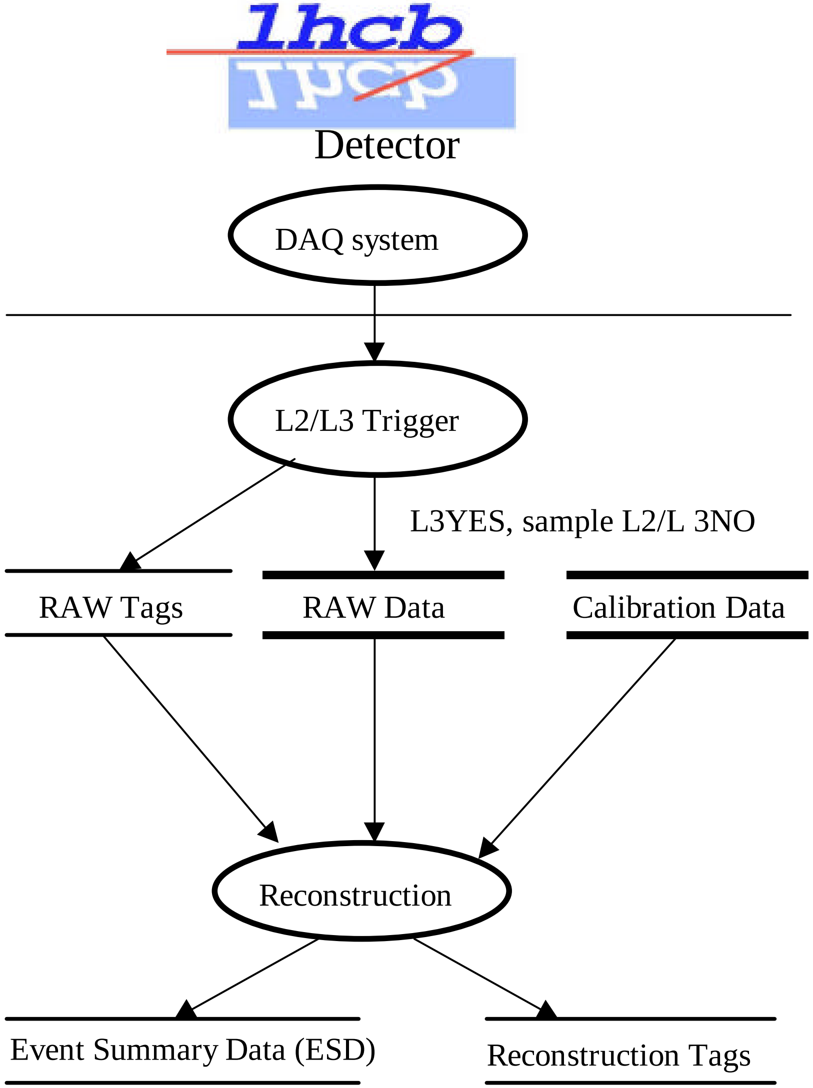
- LHCb Technical Note in preparation - see draft
- Only rough estimates of requirements are available
- Baseline model reflects current thinking
  - based on what seems most appropriate technically
  - discussions are just starting
- Open meeting of the collaboration April 5-7
  - feedback and changes can be expected

Physics Goals

- Study CP violation by measuring different final states of rare B-meson decays (>20 channels) e.g.
  - $B_d \to \pi\pi$ 6.9 kevents / year
  - $B_d \to K\pi$ 33k events / year
  - $B_d \to J/\Psi K_s$ 56k events / year
  - $B_d \to D^* \pi$ 800 k events / year
  - $B_s \to J/\Psi \phi$ 44 k events /year
  - ...
- About 10^{12} bb pairs produced in 1 year
- Numbers of interesting events reconstructed offline varies according to channel (10^5 to a few hundred)
- Trigger on high P_t and displaced secondary vertices

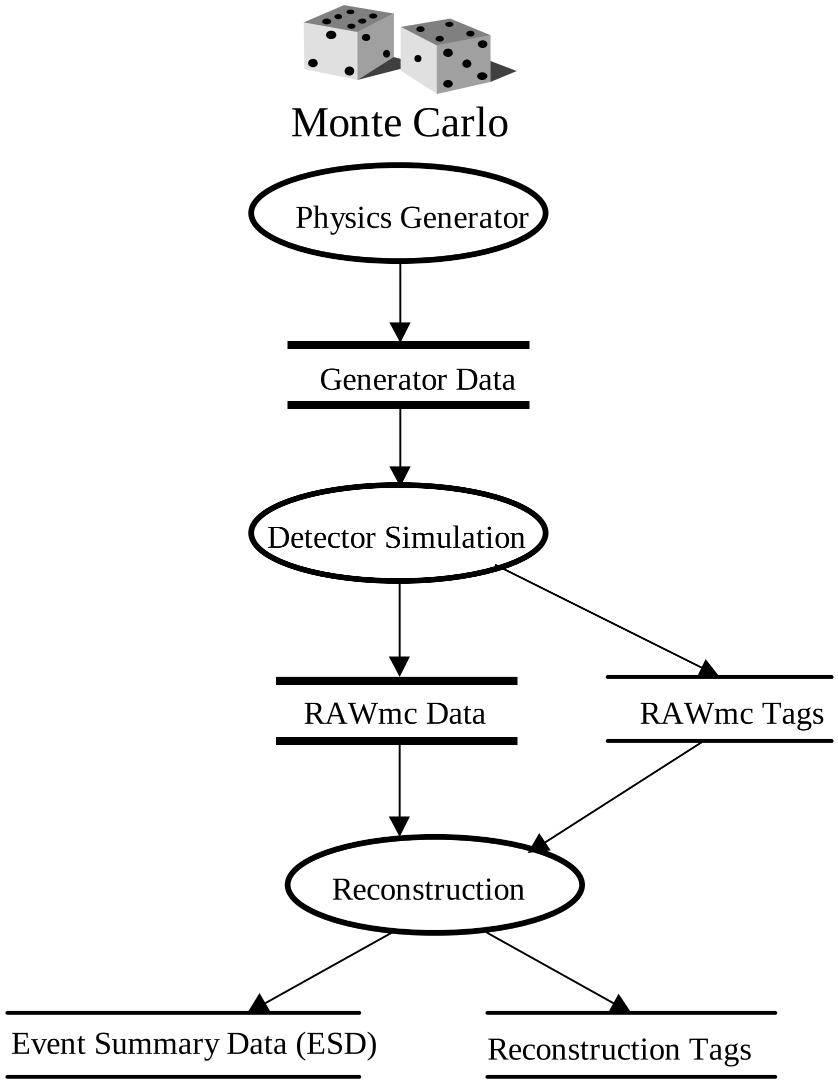
Dataflow Model - Production

- L2/L3 runs algorithms use partial reconstruction of final states
  - Detailed studies still to be made
- Small samples of rejected events kept for efficiency studies
- Reconstruction determines raw physical quantities such as energy in calorimeter, assigns hits to tracks etc.
- Reconstruction is repeated a number of times (~2) to accommodate changes in algorithms, calibration and alignment



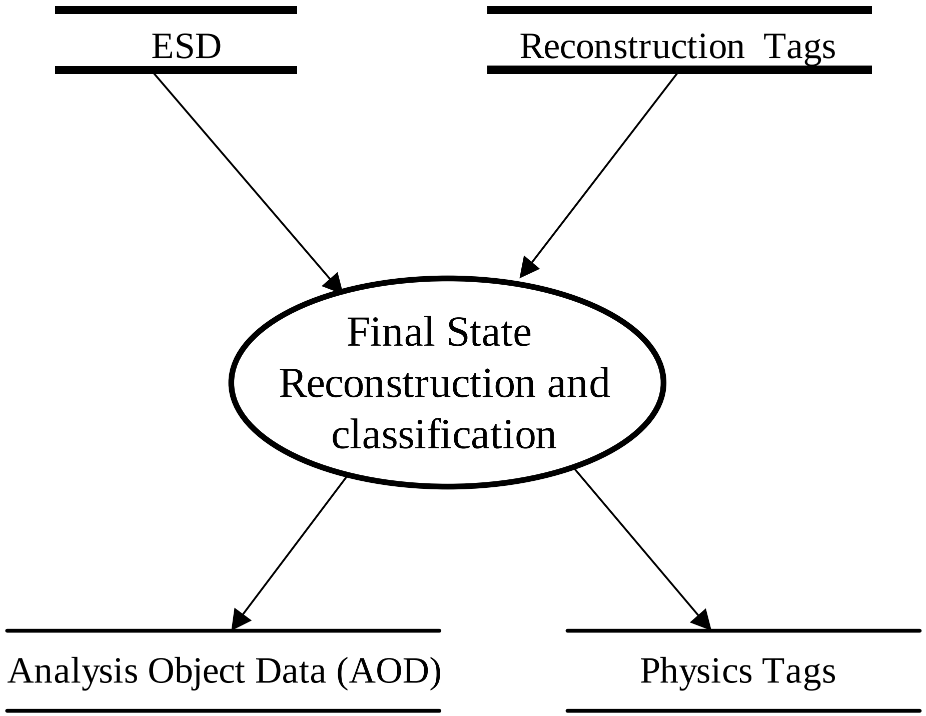
Event Simulation

- Simulates all steps :
  - bb generation
  - GEANT tracking
  - Digitisation
  - trigger
  - reconstruction
- Truth information is stored to record physics history of the event
  - RAWmc larger than RAW
- Simulation also repeated as algorithms evolve , and as detector design continues to be optimised



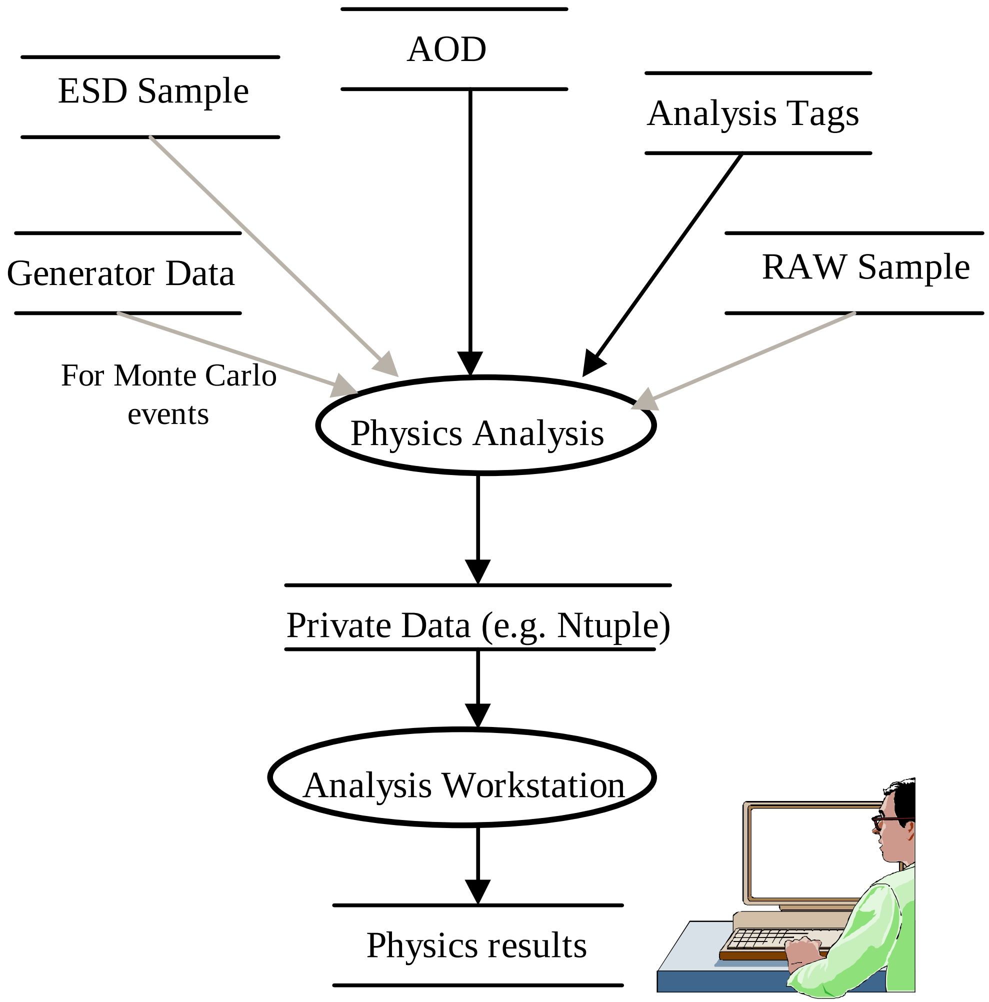
Dataflow Model - Final State Reconstruction

- Determine P^{μ} of measured particle tracks, vertices, invariant masses
- Run tagging algorithms to identify candidates for composite particles (J/Ψ, π^0)
- Common to different decay modes - run in production as First Pass Analysis
- Use Reconstruction Tags to optimise
- Step will be repeated (3-4 times/year) on complete sample as algorithms evolve



Dataflow Model - User Analysis Cycle

- Physicist runs physics analysis jobs
- Select interesting events using tags
- Reconstruct B-decay channels of interest using AOD only
  - copy parts of ESD if needed
- Generate private data (e.g. ntuple)
- Study systematic effects by looking at ESD for small event samples
- Access raw data of individual events (~100) and study in detail e.g. with event display



Real Data Processing Requirements

Length of period	120 days	10^7 secs	
LHC duty cycle	50%		
Event rate stored	200 Hz	10^7 per day	10^9 per year
RAW data size	100 kB/event	1 TB/day	100 TB/yr
ESD data size	100 kB/event	1 TB/day	100 TB/yr
AOD data size	20 kB/event	0.2 TB/day	20 TB/yr
TAG data size	1 kB/event	0.01 TB/day	1 TB/yr
L2 trigger CPU	0.25 SI95sec/event	@40 kHz	10,000 SI95
L3 trigger CPU	5 SI95sec/event	@5 kHz	25,000 SI95
Reconstruction CPU	250 SI95sec/event	@200 Hz	50,000 SI95
First Pass Analysis	5 SI95/event	2.10^8 in 2 days	5000 SI95
User analysis at RC	20 SI95/event		10,000 SI95
User analysis CERN	20 SI95/event		20,000 SI95

User Analysis Requirements

- Assume that physicist performs a production analysis and requires a response time of 4 hours
- The ~10^7 events tagged by first pass analysis are scanned and candidates selected (0.25 SI95 /event)
- The selected candidates are subjected to analysis algorithm (20 SI95 / event)
- Total installed cpu power needed calculated assuming:
  - ~140 physicists actively doing analysis
  - each submits 1 job / day (NB. many short jobs as well)
  - analysis distributed over a number of regional centres (~5) and assume ~20 physicists at each Regional Centre, ~40 at CERN
  - Assume 0.3 x 10^7 events selected for algorithm on average
  - 10,000 SI95 at each Regional Centre, 20,000 SI95 at CERN

Simulation Requirements - Signal Events

- CPU power to simulate 10^7 B -> D*π events in 1 year
  - assume need to simulate 10 times real data sample (10^6)
  - N.B.this channel dominates
  - installed capacity needed is 100,000 SI95

Step	Number of events	Cpu time/evt	Total cpu power
Generator	10^{10}	200 SI95sec	2. 10^{12} SI95sec
GEANT tracking	10^9	1000 SI95sec	10^{12} SI95sec
Digitisation	10^9	100 SI95sec	10^{11} SI95sec
Trigger	10^9	100 SI95sec	10^{11} SI95sec
Reconstruction	10^8	250 SI95sec	2.5 10^{10} SI95sec
First Pass analysis	10^7	20 SI95sec	2. 10^8 SI95sec

Simulation Requirements - Background

- 10^5 $b\bar{b}$ inclusive events in detector every second
- ~100 Hz are recorded in real data
  - trigger efficiency 10^{-3}
- If need as many to be simulated then need to generate, track, digitise and trigger 10^{12} bb inclusive events/yr and 10^9 will have to be reconstructed
  - corresponds to 3. 10^{14} SI95 sec / yr
- Obviously need to study ways of optimising background simulation
  - store and reuse data produced at generator level
  - optimise generation step without biasing physics
  - focus on background particularly dangerous for a specific physics channel
  - aim to reduce requirements by > 1 order of magnitude
- Assume 400,000 SI95 required

Simulation Requirements - Summary

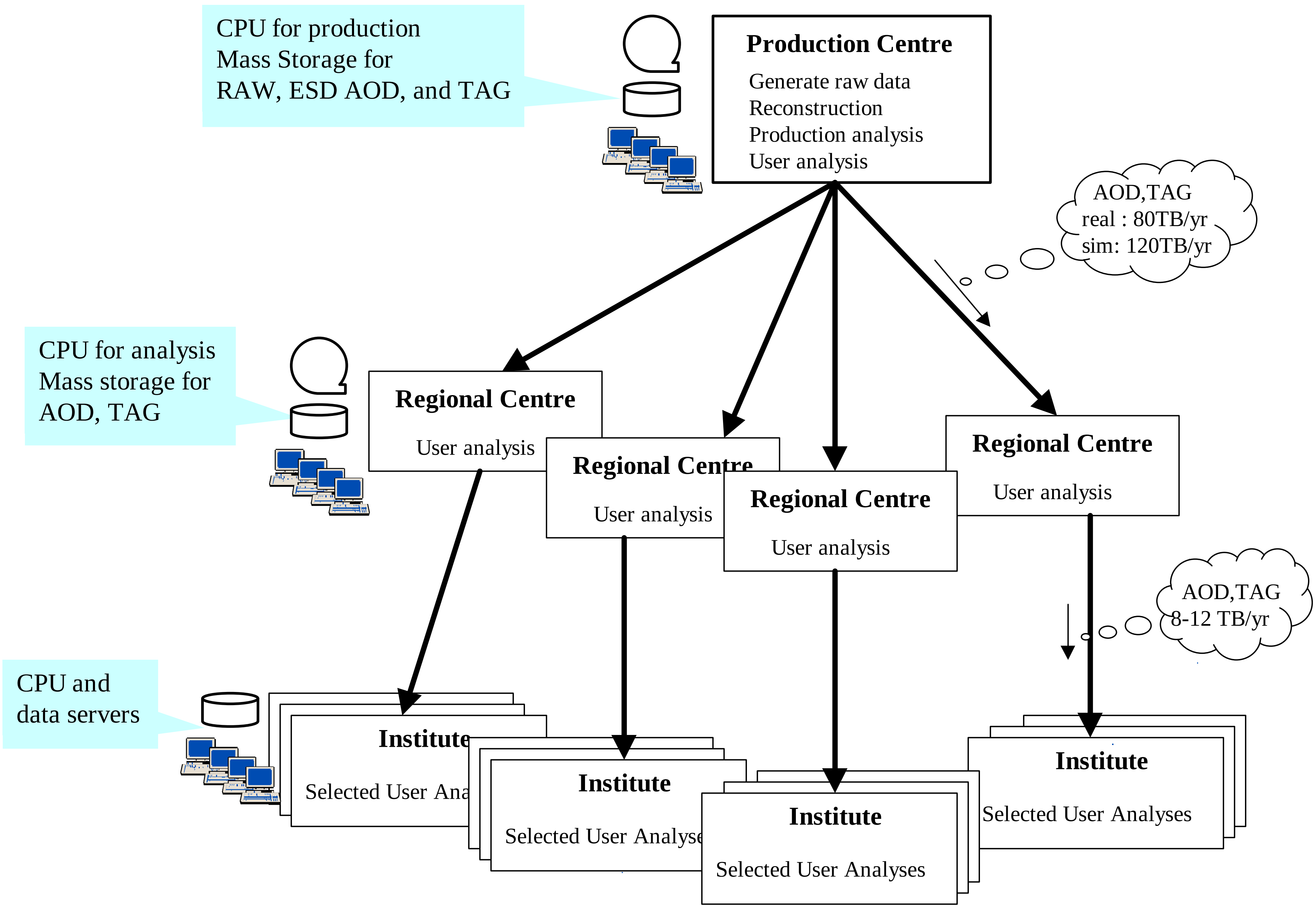
RAWmc data size	200 kB/event	200 TB/10^9 events
Generator data size	12 kB/event	12 TB/10^9 events
ESD data size	100 kB	100 TB/10^9 events
AOD data size	20 kB/event	20TB/10^9 events
TAG data size	1 kB/event	1 TB/10^9 events
CPU power	~100,000 SI95 signal events	~400,000 SI95 background events

Baseline Computing Model

- ❑ Based on a distributed multi-tier regional centre model
- ❑ Identify the production centre
  - ➢ responsible for all production processing phases
  - ➢ generation, reconstruction, and first pass analysis
- ❑ Production Centre archives all data generated
  - ➢ RAW, ESD, AOD, TAG (+ generator for simulation)
- ❑ Assume bulk physicis analysis normally only requires access to AOD and TAG datasets
  - ➢ specific ESD data needed in analysis (small) added to AOD
  - ➢ only ship AOD and TAG outside to other centres
- ❑ Assume analysis repeated ~4 times per year
  - ➢ 80 TB / yr (real), 120 TB/yr (simulated) to each RC
  - ➢ Move data using most appropriate medium (network, tape...)

Baseline Computing Model - Roles

- To provide an equitable sharing of the total computing load can envisage a scheme such as the following
- After 2005 role of CERN
  - to be production centre for real data
  - support physics analysis of real and simulated data by CERN based physicists
- Role of regional centres
  - to be production centre for simulation
  - to support physics analysis of real and simulated data by local physicists
- Institutes with sufficient cpu capacity share simulation load with data archive at nearest regional centre
- NB This scheme still to be discussed in the collaboration - political issue as well as technical


CPU for production
Mass Storage for
RAW, ESD AOD, and TAG
Production Centre
Generate raw data
Reconstruction
Production analysis
User analysis
AOD,TAG
real : 80TB/yr
sim: 120TB/yr
CPU for analysis
Mass storage for
AOD, TAG
Regional Centre
User analysis
Regional Centre
User analysis
Regional Centre
User analysis
Regional Centre
User analysis
AOD,TAG
8-12 TB/yr
CPU and
data servers
Institute
Selected User Ana
Institute
Selected User Analyse
Institute
Selected User Analyses
Institute
Selected User Analyses

	Real Data	Simulated Data
	CERN	RAL , Lyon, ...
Production Centre (x1)	Data collection Triggering Reconstruction Final State Reconstruction	Event Generation GEANT tracking Reconstruction Final State Reconstruction
↓	***WAN Output to each RC:*** *AOD and TAG datasets* *20TB x 4 times/yr= 80TB/yr*	***WAN Output to each RC:*** *AOD, Generator and TAG datasets* *30TB x 4 times/yr= 120TB/yr*
Regional Centre (~x5)	User Analysis	User Analysis
↓	***WAN Output to each Institute:*** *AOD and TAG for samples* *1TB x 10 times/yr= 10TB/yr*	***WAN Output to each institute:*** *AOD and TAG for samples* *3TB x 10 times/yr= 30TB/yr*
Institute (~x50)	Selected User Analysis	Selected User Analysis

Analysis Scenarios

- Technical Note describes three typical scenarios for a distributed physics analysis
  - physics channel under study
  - regional centres and institutes involved
  - how data are achived, and requirements on shipping to remote sites

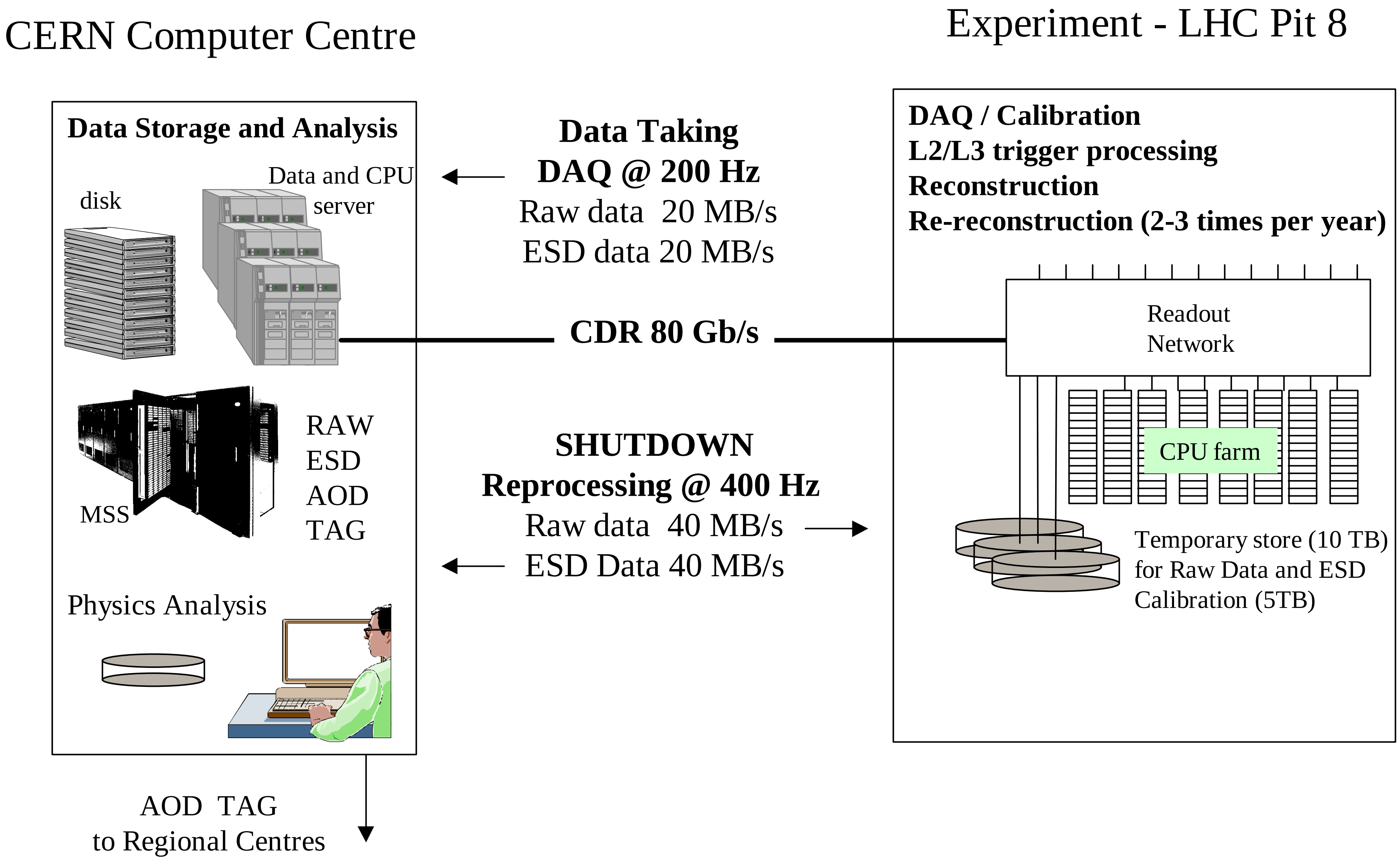
Differences with MONARC

- Do not explicitly identify Group Analyses
  - First pass analysis run as one production job as much of final state reconstruction is common to different decay channels
- Run all production data processing from RAW to AOD at the production centre. Only ship AOD and TAG datasets to regional facilities.
- Envisage to focus activities at CERN after 2005 on processing of real data and simulation at remote facilities
- Our data processing requirements do not imply a clear need to distinguish Tier 1 and Tier 2 centres.

Computing at CERN

- Run high level triggers and reconstruction on same cpu farm located at LHCb pit
- Send RAW and ESD data over the CDR link (80 Gbps) to computer centre for archive, and first pass analysis
- Maintain local storage at pit in case CDR down
  - accumulate 2 TB/day, therefore need >10 TB
- Dispatch AOD and TAG etc to regional centres
- During shutdowns, down periods do re-processing of RAW data on the farm in the pit
  - read RAW back from computer centre
  - send new ESD from pit to computer centre
  - full farm available so proceeds at twice the rate
  - allows 2-3 reprocessings of complete year's data
- Flexible, efficient and maintainable solution

Compute Facilities at CERN



Facility at Pit - Requirements

CPU Farm	~100,000 SI95
Disk storage event buffer	> 10 TB
Disk storage calibration and secondary data	> 5TB
CDR link capacity (80 Gb/s)	1 Gb/s

CERN Computer Centre Requirements

RAW data storage	100 TB/yr
Copy RAW data storage	100 TB/yr
ESD data storage	100 TB/yr
AOD data storage	4 x 20 TB/yr
TAG data storage	1 TB/yr
AODmc, Generator storage	120 TB (30 TB imported 4 times/yr)
TAGmc data storage	4 TB (1 TB imported 4 times/yr)
Total data storage	~500 TB / yr
CPU for First Pass analysis	2000 SI95
CPU for user analysis	20,000 SI95
WAN for AOD TAG export	80 TB/yr
WAN for AOD TAG import	124 TB/yr

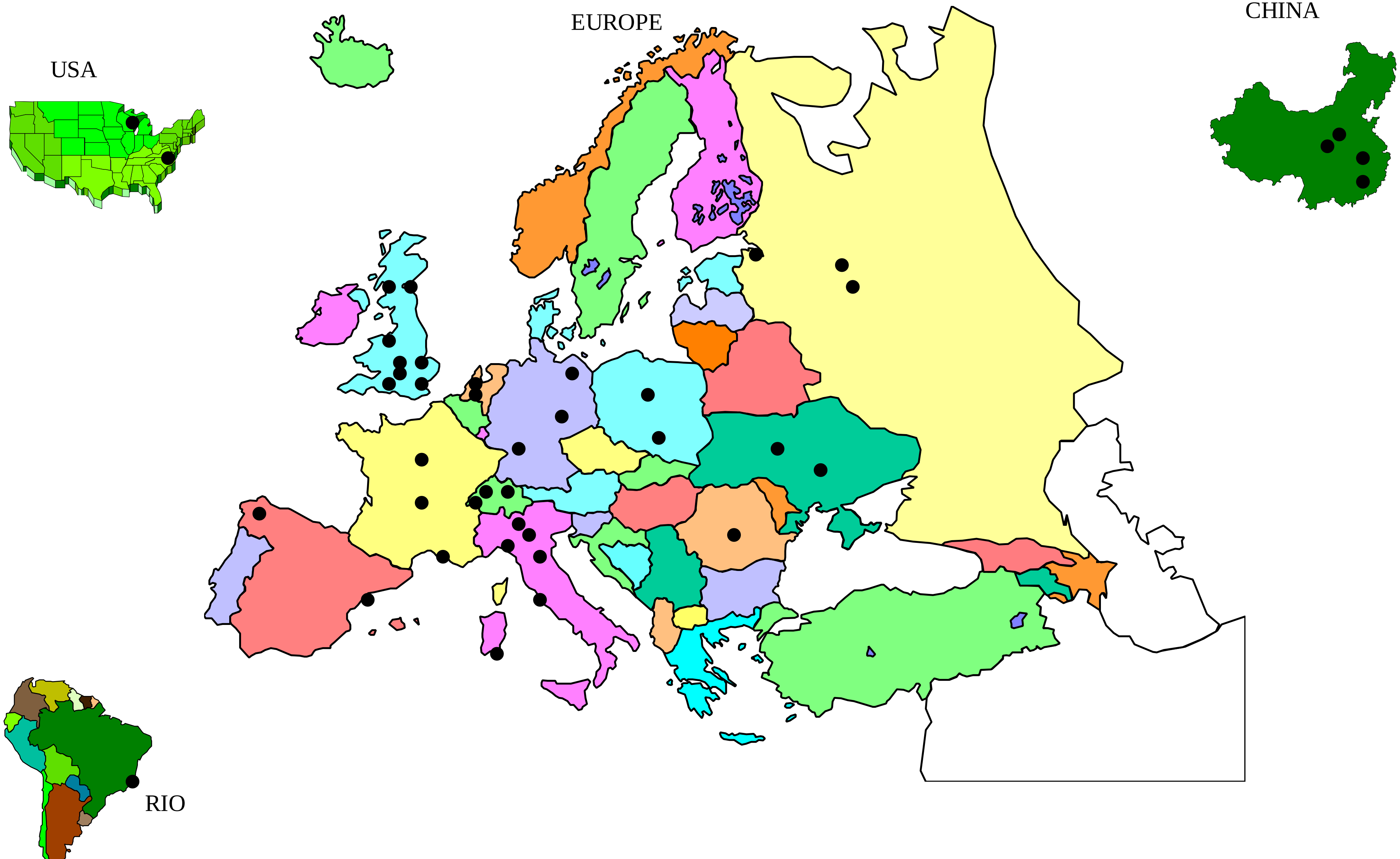
Simulation requirements 2000-2005

- 2000-2001
  - 10^7 simulated events/yr for detector optimisation studies
  - prepare TDRs
- 2002-2003
  - $2. 10^7$ events/yr for high level trigger studies
- 2004 - 2005
  - start to install and commission large scale facilities
  - start to produce large samples of background events with the final detector description
  - ~10^8 simulated events/yr
- >2001
  - use simulation and MDC to test computing model
  - contribute to HEP Application WP of EU grid proposal (scenario3)

Sizing Estimate for Regional Centre

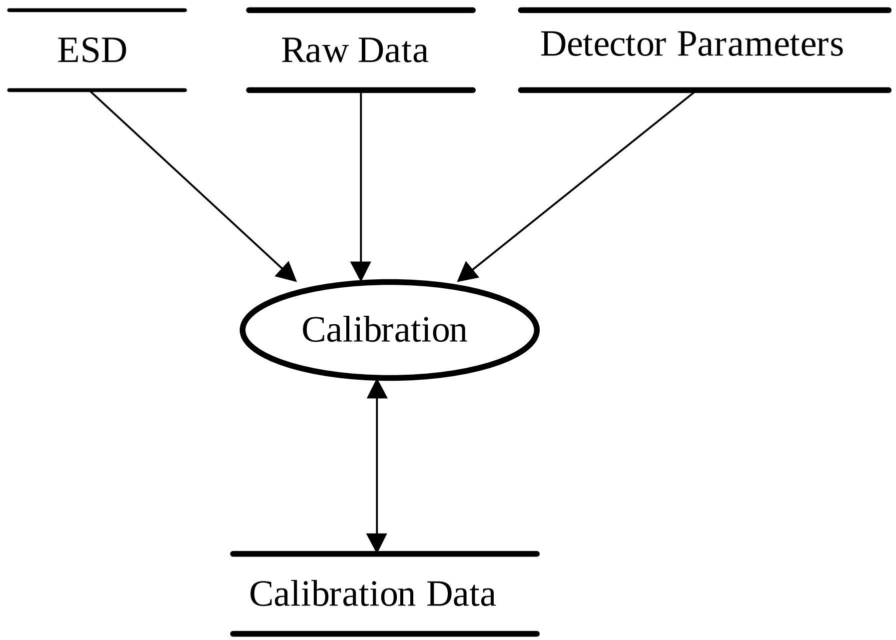
	2000-2001	2002-2003	2004-2005	>2005
AOD TAG				80TB/yr
AODmc TAGmc imported	2TB/yr	5TB/yr	20TB/yr	120 TB/yr
CPU analysis	3000 SI95	5000 SI95	10000 SI95	10000 SI95
RAWmc, ESDmc AODmc TAGmc generated	5TB/yr	10TB/yr	33TB/yr	333TB
CPU mc production	20000 SI95	40000 SI95	60000 SI95	100000 SI95

LHCb Collaboration Institutes



LHCb Dataflow Model - Calibration

Calibration Cycle



- IO overhead (20% from BaBar)