HLT Large Scale Tests 2006
and HLT Installation at Point1

At the LXBATCH facility at Cern/IT,
farm size increasing throughout the testing period up to 1000 nodes

Goals

- ***Configure HLT** infrastructure and algorithms, both Level 2 and Event Filter*
- ***Exercise DB access** for configuration, in particular*
  - *DB caching for the HLT*
  - *the TriggerDB*

  } *both new, essential ingredients*
- *Run **HLT with a set of trigger slices and algorithms** on a realistic mix of MC events*
- ***Configure** (emulated) **ROD crates***
  - *with DBStressor developed for LST06*
- *Infrastructure timings, monitoring*
  - *detailed tests already performed in LST05*
  - ***verify the good results** in LST06*

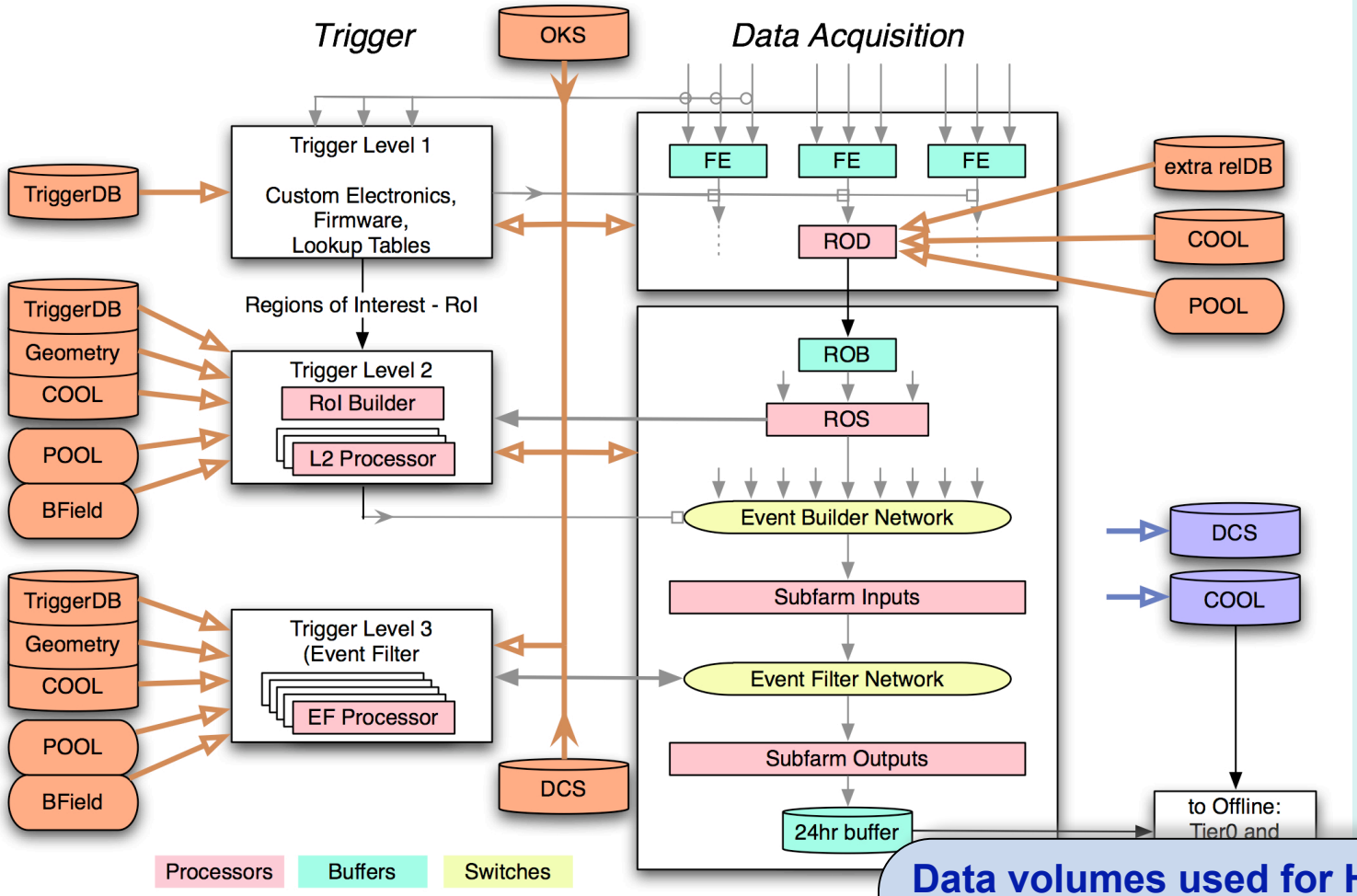
- November/December 2006
- dual cpu pentium IV 2.4-2.8 GHz nodes; slc4
- inhomogeneous network
- dedicated MySQL servers
- Set up and maintained by IT

Haimo Zobernig, Serge Sushkov, Hans von der Schmitt,
Sarah Demers, Jörg Stelzer, Doris Burckhart, Lourenco Vaz, Sergei Kolos, Haleh Hadavand

Databases involved at LST06

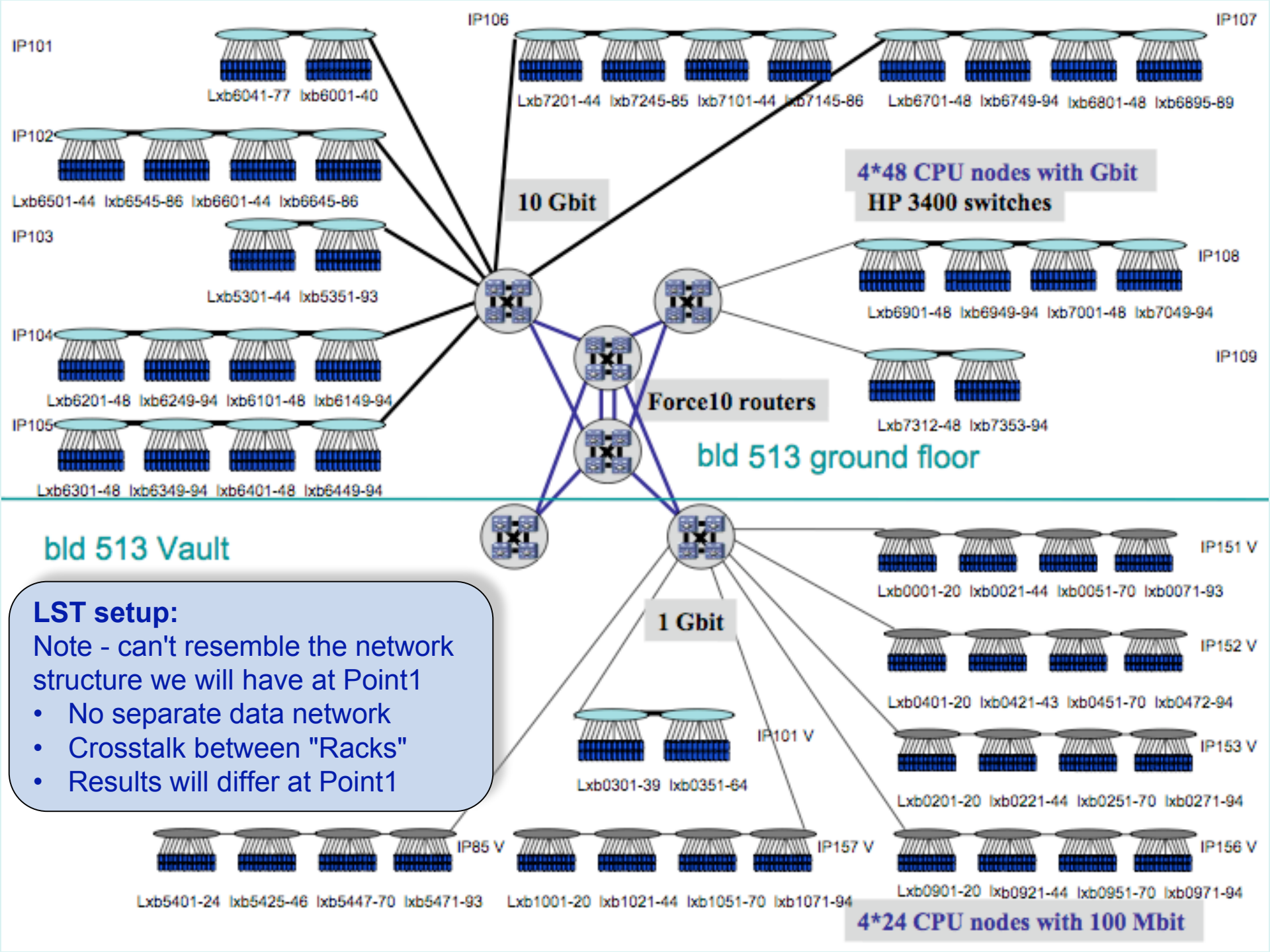
- ***Configuration using various databases with these options:***
  - ***OKS to define the partition***
  - ***DCS not used in the LST06***
  - ***COOL (via CORAL): Oracle, MySQL, DbProxy, SQLite***
  - ***Geometry (via CORAL): SQLite (because it is ~invariant)***
  - ***TriggerDB (via CORAL): Oracle, MySQL, DbProxy; as well as job options***
- ***On the local filesystem of each node we had:***
  - ***The software (TDAQ-01-06-02, 12.0.3-LST, HLT-2-0-3)***
  - ***SQLite, POOL, and BField files***
- ***Oracle server: online cluster ATONR***
  - ***2 server nodes - now upgraded to 6 nodes***
  - ***Located at IT, still on CERN public net - soon on ATCN***
  - ***Oracle is the primary database resource for real running (configuration, DCS, conditions) - was not yet the case in LST06***
- ***MySQL servers:***
  - ***2 nodes installed on the LXSHARE cluster***

Configuring the full system from DBs



Data volumes used for HLT configuration:

- Conditions database 38 MB
- Geometry database 23 MB (SQLite file)
- Trigger configuration database 1 MB
- POOL files 31 MB including the catalogues
- Magnetic field file still unrealistic 5 MB.

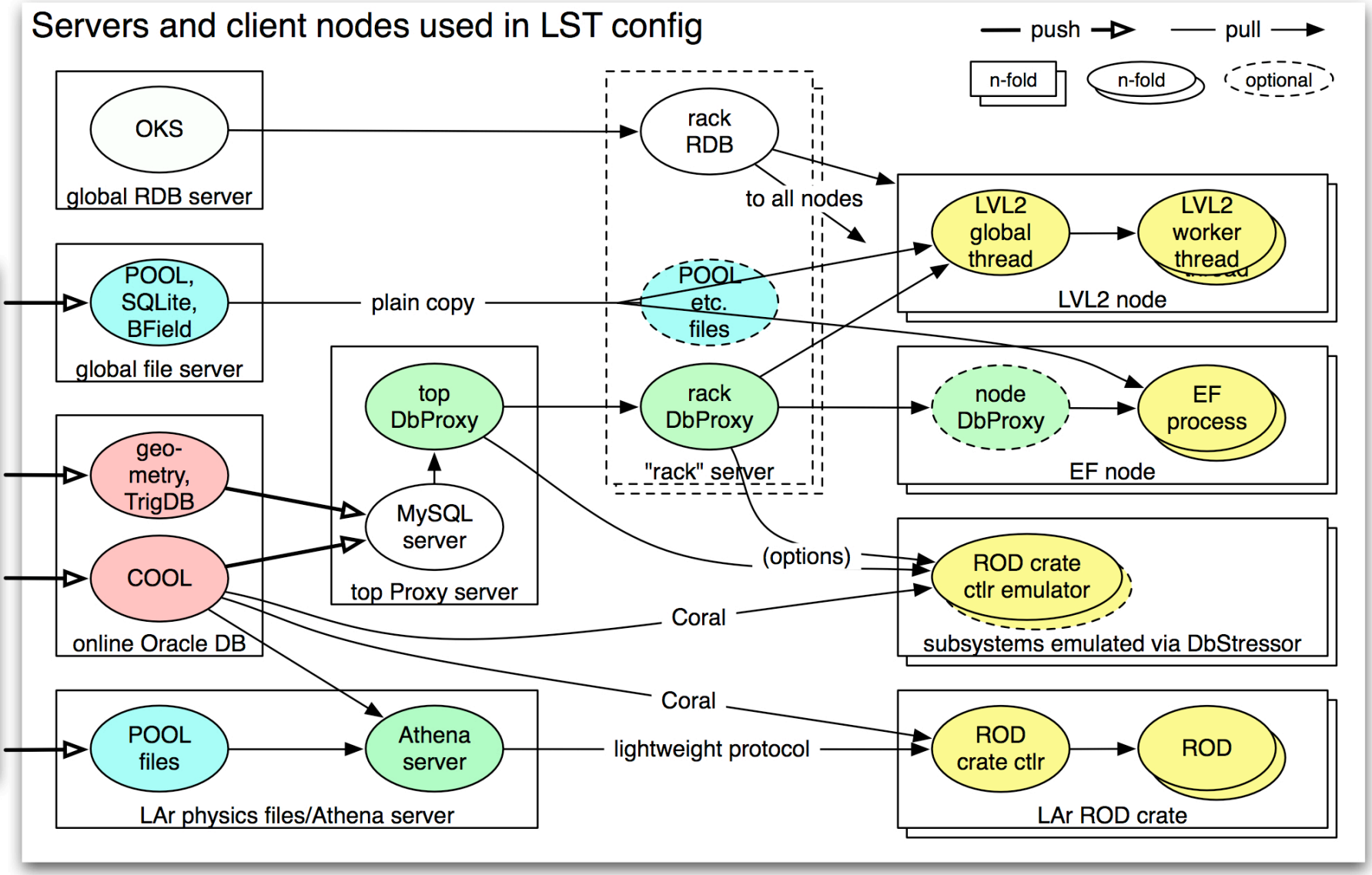


LST setup:

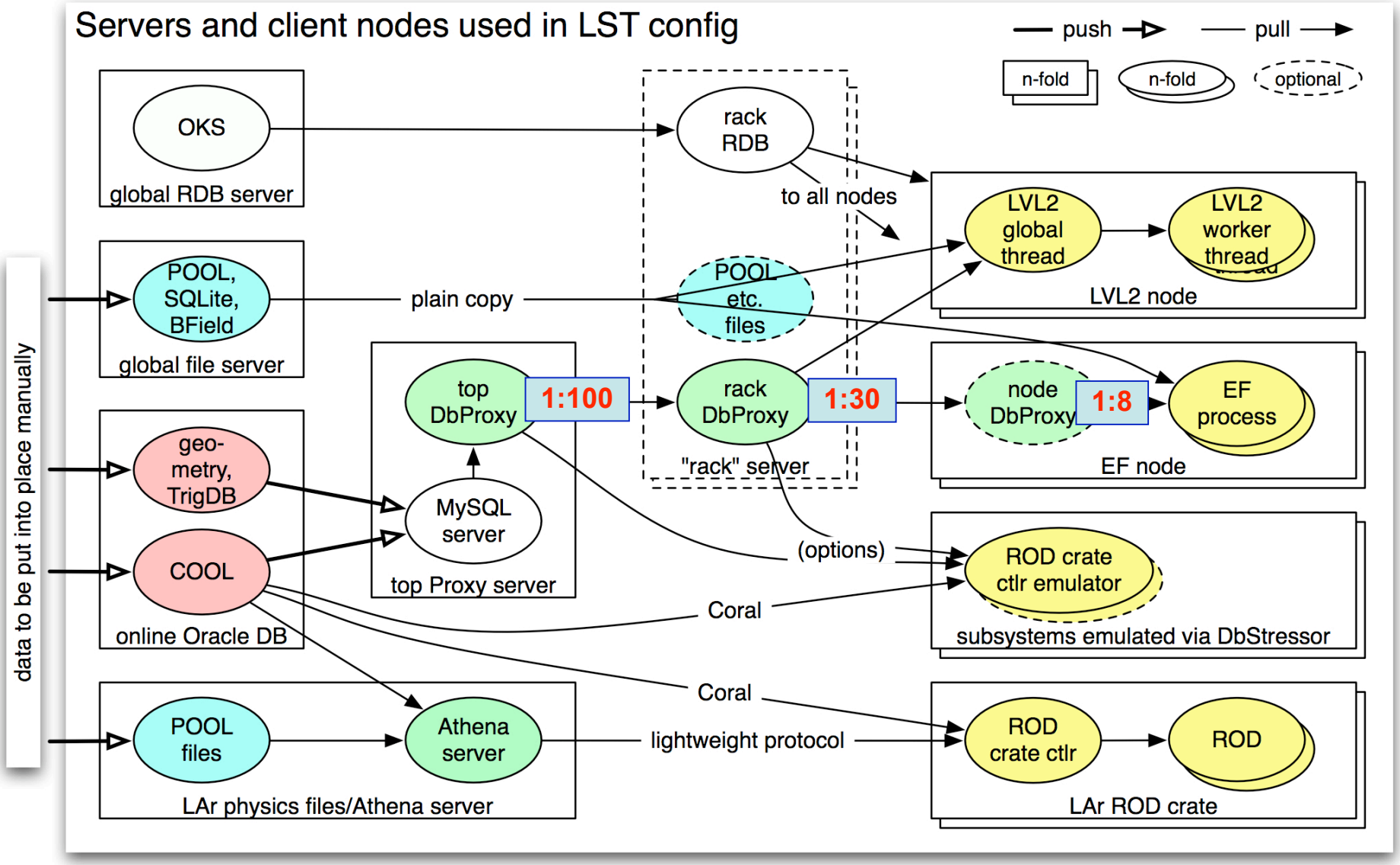
Note - can't resemble the network structure we will have at Point1

- No separate data network
- Crosstalk between "Racks"
- Results will differ at Point1

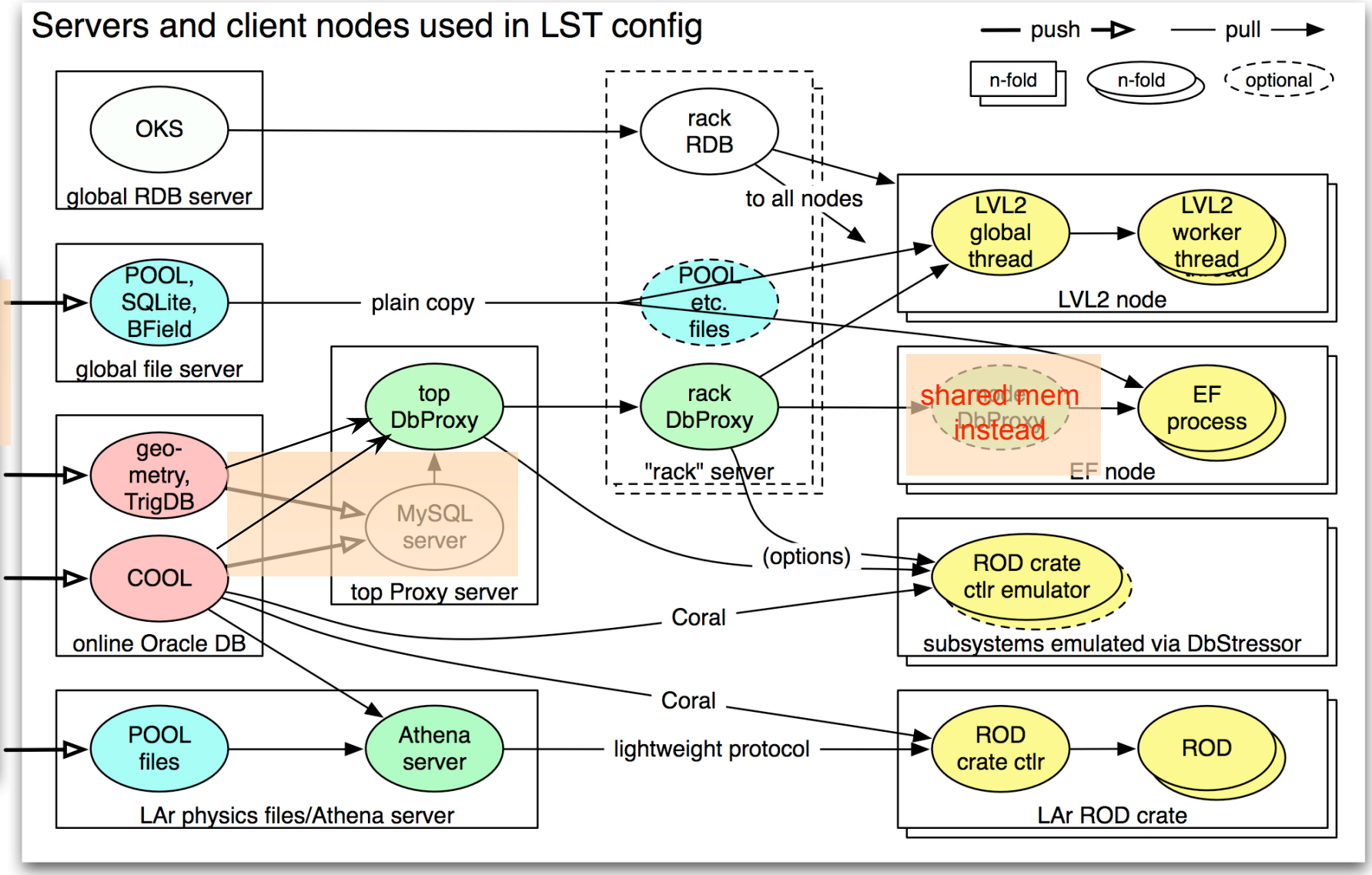
DB servers, file servers, and clients in LST



DB servers, file servers, and clients in LST



DB servers - ongoing development



Level 2 and Event Filter tests

- ***Objectives:***
  - ***Run large DAQ partitions; Run with DbProxy, reading CondDB (COOL data)***
  - ***Study L2/EF with algorithms and CondDB configure times at various scales and SubFarm sizes and modes of DbProxy use***
- ***Level 2:***
  12 SubFarms, up to 40 L2 Processing Units per SubFarm = 480 L2PUs, no Event Builder

  - Algorithms configured from TriggerDB or jobOption files
  - Transition time for each L2PU measured individually
  - Geometry from SQLite DB file local on each node
  - POOL files and B field map local on each node
  - Conditions data from SQLite, MySQL, Oracle, or DbProxy

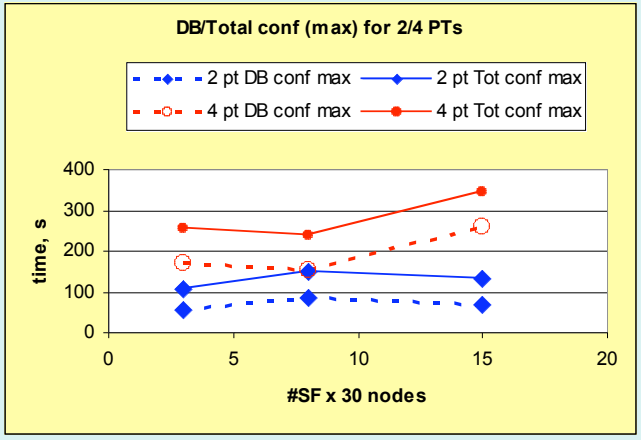
- ***Configure***
  - ***L2 configuration times quite acceptable: ~1.5 min - much faster than 2005***
  - ***Don't depend strongly on where configuration data come from: MySQL, Oracle, SQLite files, DbProxy cache***
  - ***Configure time is not dominated by DB access (which is ~15 sec)***
- ***Stop***
  - ***Many/most L2PUs stop within a few seconds***
  - ***"slow stoppers" of several min due to events with excessive time in HLT algos***

Level 2 and Event Filter tests

- ***Event Filter***
  measurements with various SubFarm configurations with 450 Event Filter Dataflow Tasks (EFD), and up to 1800 Processing Tasks (PT)

  - Focus on HLT-CondDB-DbProxy performance
  - Varying configuration sizes of EF trigger algorithm
  - Measure Conditions DB access time as part of configure transition
  - Algorithms configured from jobOption files
  - COOL data access directly or via SQLite, MySQL, DbProxy
  - Transition times for PTs measured individually

Scaling of total DB access times with size of EF System: 3, 8, 15 SubFarms of 30 nodes each, 2 and 4 PTs/node



- ***Configure***
  - ***EF configuration time: ~2 min***
  - ***Scalability exercise (2PT/CPU): > 6 mins with overloaded CPU***
  - ***topology of farm layout will make a difference – better at Point1***

DbProxy

- *Example of one DbProxy serving 20 nodes i.e. 40 L2PUs*
- *Each L2PU during configure makes 6 DB transactions i.e. 6 connects, 6 disconnects, plus the queries inbetween*

- *Upper histogram:*

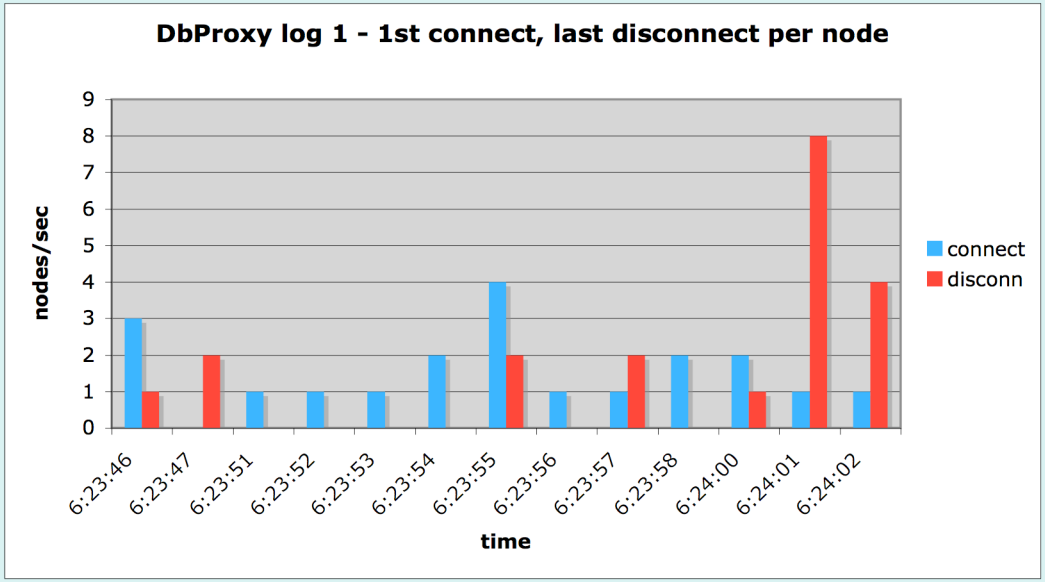
  absolute time of 1st connect and of last disconnect, one entry per node

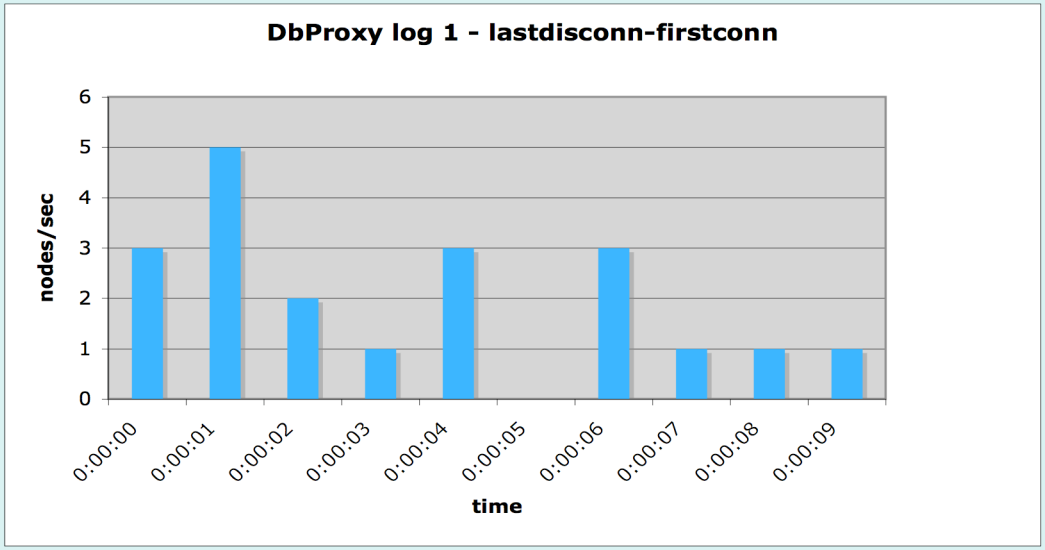
  total span 17 seconds

- *Lower histogram:*

  time difference last disconnect minus 1st connect, one entry per node

  maximum duration 9 seconds





Conclusions and Outlook

- ***LST06:***
  first time the DAQ/HLTsystem was run successfully with full configuration from databases, and at large scale
  - ***Separate tests for LVL2 and for EF***
  - ***Configuration timing reasonable and apparently not dominated by database access***
  - ***A number of bugs found - fixed "online" or immediately after the LST***

- ***The complete DAQ/HLT system with ROS, LVL2, Event Builder and EF without algorithms was run on up to 600 nodes***
  - ***Monitoring and Run Control timings OK***
  - ***Need to work on Fault Tolerance throughout the system***
  - ***Still improvements necessary in the area of Farm Tools, PartitionMaker***
  - ***Operational Monitoring Tools to be put in place***

- ***Thanks to IT for the valuable support***

- *A full report was given at the TDAQ open meeting 8. February 2007*
  for the slides see *http://indico.cern.ch/conferenceDisplay.py?confId=11815*
  The written report is being finalized on Twiki AtlasTDAQLargeScaleTests2006Report

Conclusions and Outlook (2)

- *More complete tests foreseen on the hardware installed at Point1*
  - *With the adequate network structure within racks*
  - *Using the pre-series machines, and also new 8-core nodes*
  - *With all four trigger slices configured from TriggerDB*
    - *In LST one e-gamma slice from DB - all slices only with jobOptions for Level2*
  - *Investigate details of trigger algorithm setup timing at configuration transition, esp. CPU intensive parts*
  - *Rack-specific configuration of CORAL to be integrated into a partition generation tool*
    - *DbProxy node name depends on rack - had to use inelegant scripts in LST*
  - *Further tests at SLAC (DbProxy specific) and Manchester (trigger specific)*

Status of DAQ/HLT Purchasing & Installation

- High Level Trigger nodes
  - 130 dual Quad-core machines, i.e. 8 cores in 1U
    - DEL PowerEdge 1950
      - Clovertown 1.86GHz, 1Gig./core
  - Aim to complete installation & standalone commissioning by end March
  - Another HLT rack in process of being purchased
    - Dual dual-core, i.e. 4 cores
      - WoodCrest 3.0GHz, 1Gig./core
    - Expected around end March
  - For 2008 running
    - At least 30 Racks
    - Aim to start receiving machines in Feb. '08
      - Start purchasing procedure ~ Aug. '07
    - Complete installation ~May
      - Including standalone commissioning
      - Ongoing deliveries used to define procedures & update time required

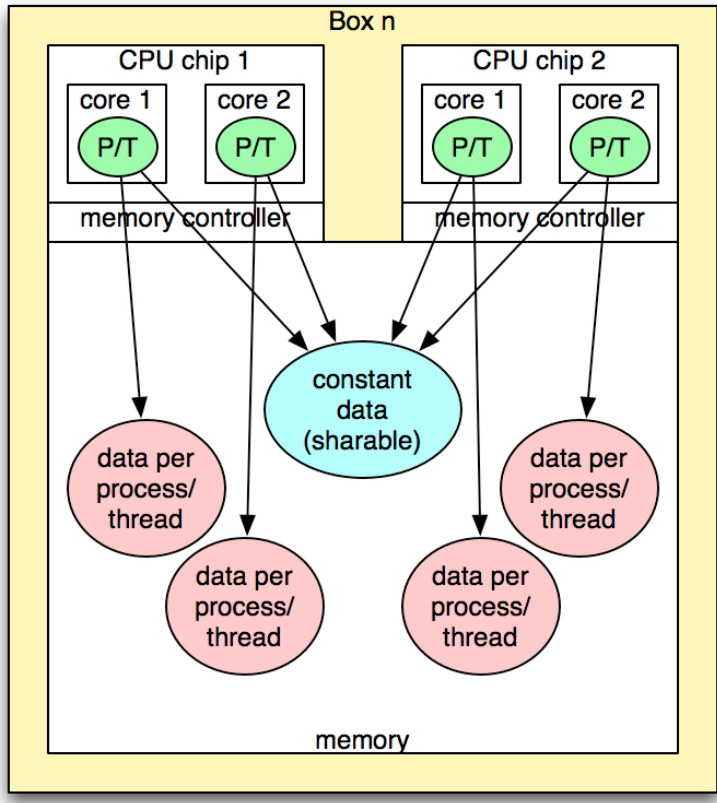
David Francis
TDAQ General Meeting 8 Feb

Conclusions and Outlook: Caching

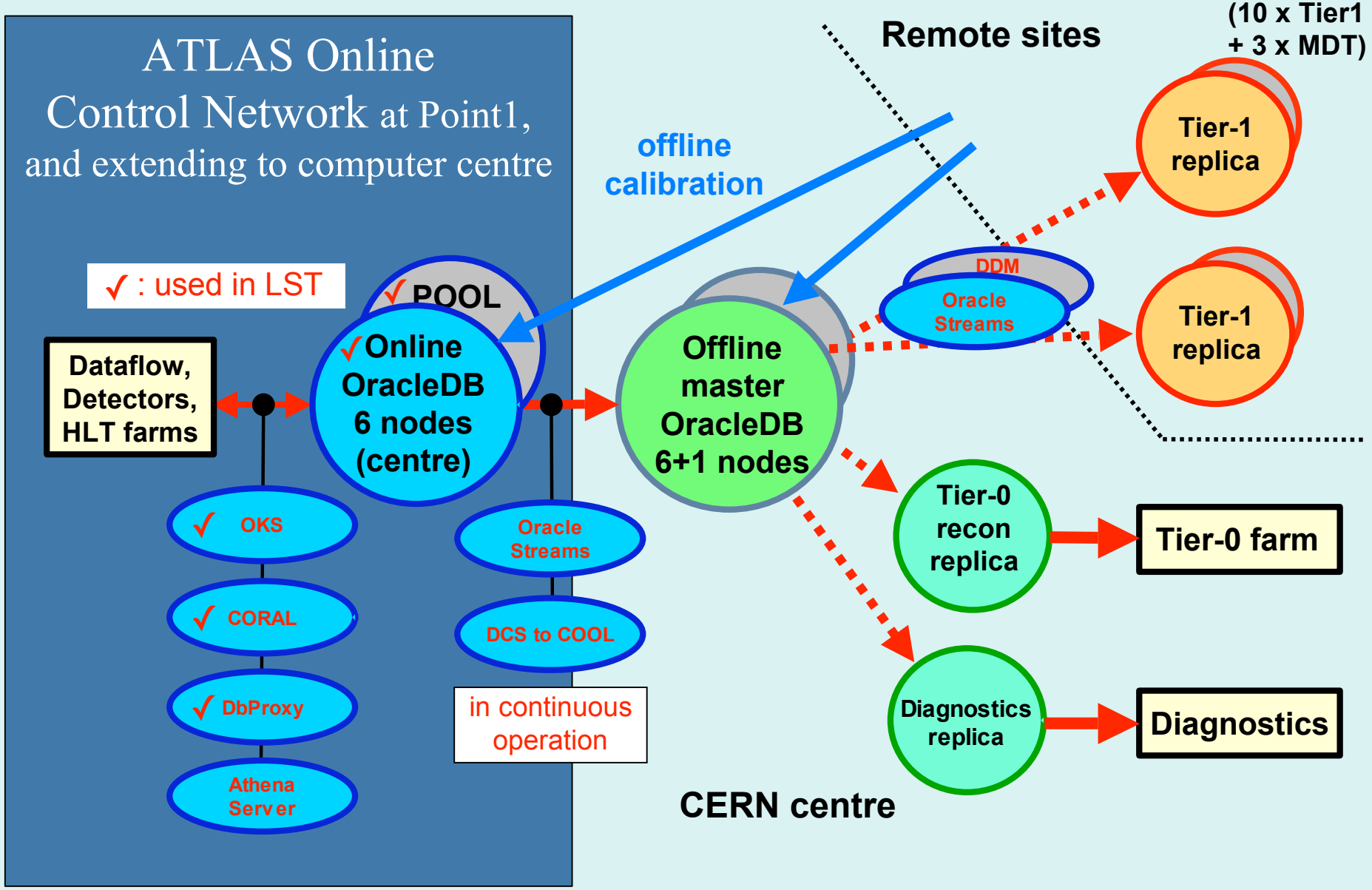
- ***HLT configuration via DBProxy caches was a success***
  - ***We have a proven way to scale from one to many HLT racks***
  - ***SLAC group now working on a direct DbProxy to Oracle connection***
  - ***DbProxy used for data varying for each run: conditions + TriggerDB***
  - ***Geometry and magnetic field map are taken from files***
- ***At node level, caching could be done as well, but...***

We have nodes with 2*4 cores each already today - more cores to come

  - ***need to optimise memory consumption and initialisation effort - options:***
  - ***global (initialization) thread + multiple event threads share constant data (geometry, fieldmap)***
  - ***initialization process + multiple event processes could shared memory for constant data***
    - ***using shared memory segment***
    - ***fork() and utilize copy-on-write***
  - ***options for sharing are now studied in Athena architecture team, for L2, EF, and possibly T0***
  - ***Athena object store (StoreGate), if in shared memory, could be loaded with ready-made object data from a file in one go to save a lot of initialization time (but vtab...)***



Overview online & offline database connections



Two major components of HLT configuration data

HLT Trigger Menu

- Defines the list of our physics triggers

Hierarchy of Chains, Signatures (steps), and TriggerElements.

- Configures the HLTSteering

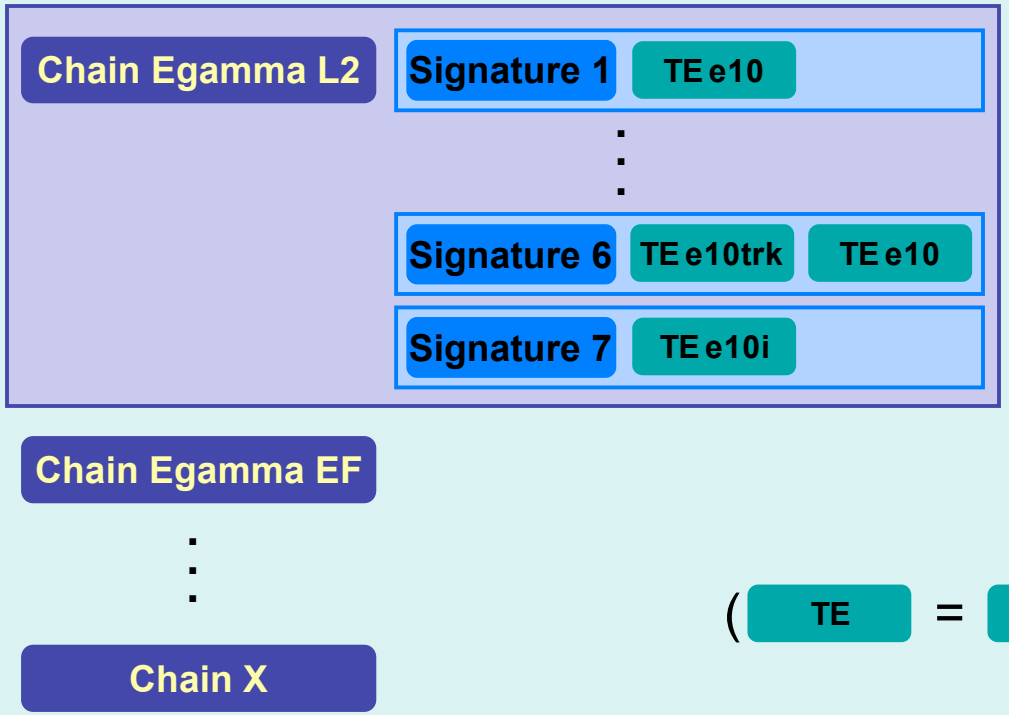
navigates the ROIs through TEs

HLT Job Parameters

- Defines physics selections

thresholds, etc., conditions (POOL references), …

- Configures all algorithms, services, and tools



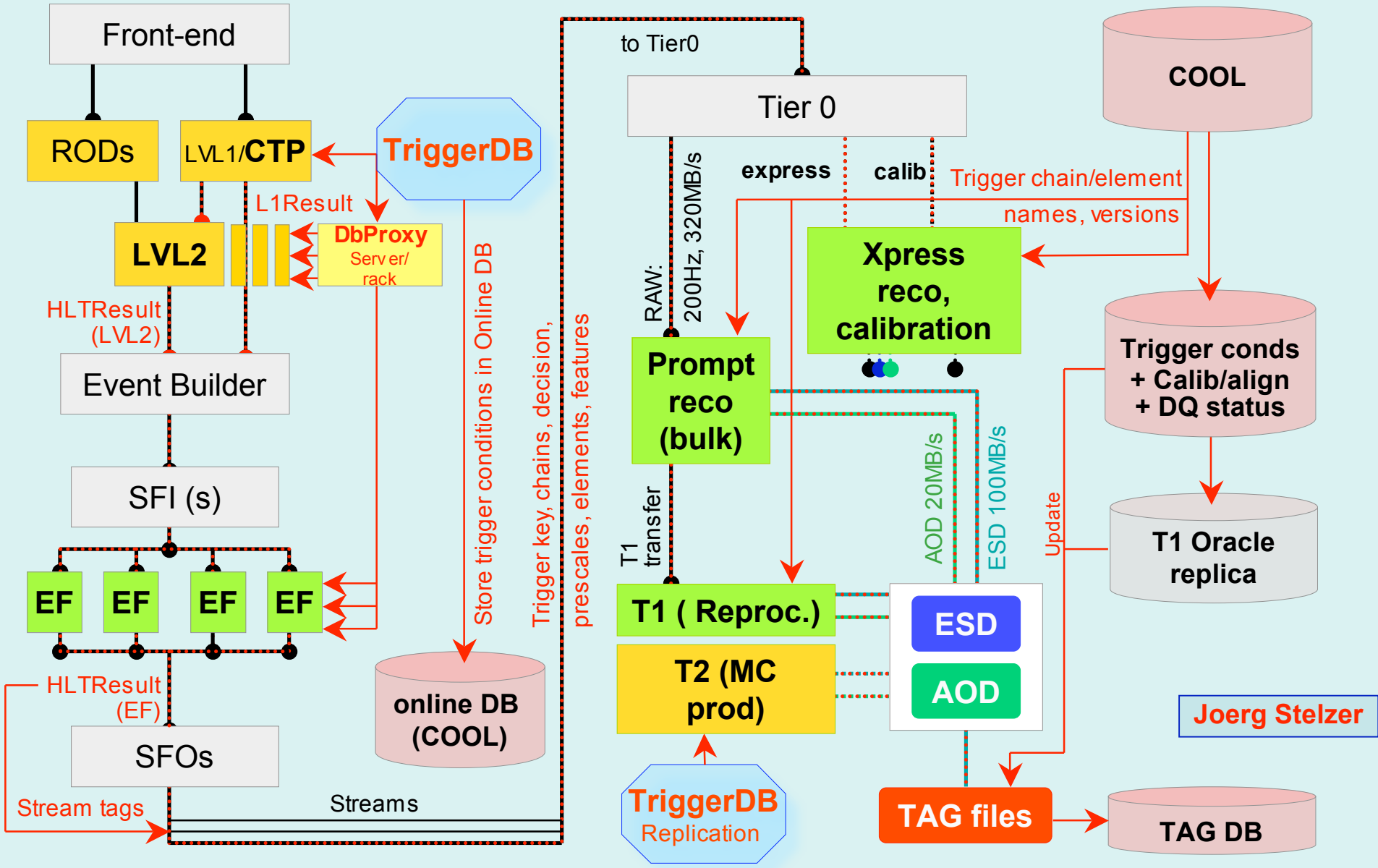

```
<Alg Name="ClusterReco"
      ClassName="ClusterReco">
 <PropName="OutputLevel" PropVal="3"/>
 .
 .
 .
 <PropName="minEpercell" PropVal="10"/>
 <PropName="separation" PropVal="15"/>
</Alg>
.
.
.
<Alg Name="…"> … </Alg>
```

(TE = HLT Algorithm)

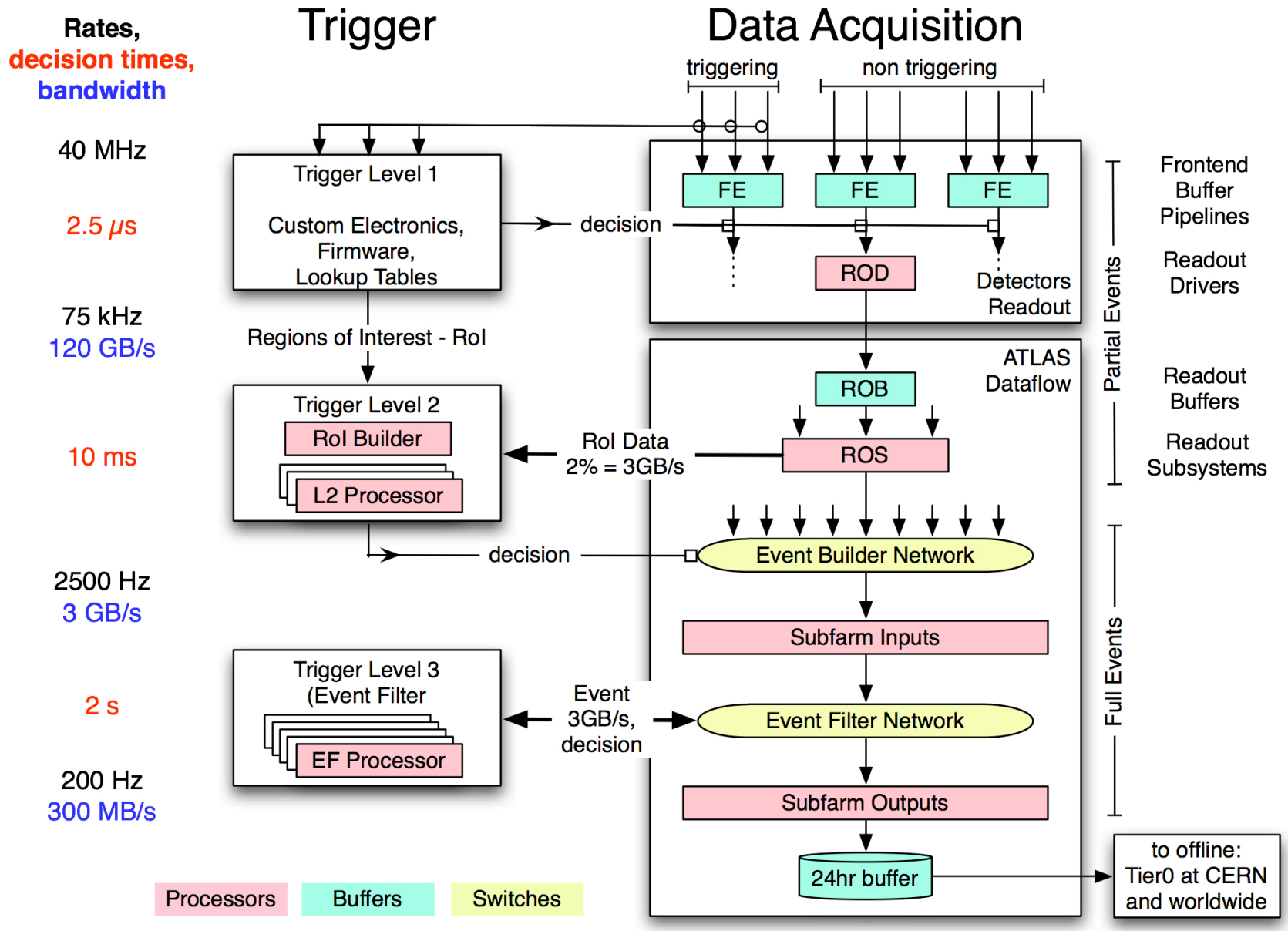
Joerg Stelzer

- *LVL 1 also configured from TriggerDB but not tested during LST*

Trigger information flow



Trigger & Data Acquisition - characteristics and acronyms



Next Software & Computing Workshop