OSG - NP experiments meeting

RHIC/STAR plans and needs …

The challenges

How to locate the **data** sets?

- **Evolution of the Scientific program** goes toward statistically challenging data samples
  - For full Phase space inspection, “bigger” everything is needed
    - Amount of data, events, files, code
    - …

How to access the (distributed) **data** sets

- **Variety and complexity of analysis** increases
  - Soon, groups cannot talk to each other due to the lack of
    - A common “language”, data format, …
    - A common interface& access method to the (shared) data
  - Others spend time to resolve the data “sorting”
    - A commodity tool addressing everyone’s needs
    - …

How to store and/or move the **data** sets

- Tight **budget** force daring (but cheap) **technology choice**
  - Surely the case for storage – concept of (local) distributed data
  - Cheap solution often lacks off-the-shelf data access method

How to evolve and keep productivity

- THE Challenge is about
  - **How to resolve all of the above and STILL allow Physicist to do their work**

The challenges

How to locate the **data** sets?

How to access the (distributed) **data** sets

How to store and/or move the **data** sets

How to evolve and keep productivity

- How to build a (data) Grid allowing
  - **Experts to run simulations, data production**
    - With under the hood data handling, registration in Catalog, …
    - Access to database
  - **A VO manager (production, data) to locate and manage datasets**
    - Invalidate some world-wide even after distribution
  - **Users to run user analysis jobs**
    - expressing their needs in a user-driven fashion
    - Accessing data in a transparent manner without the work of “data placement”
    - with the same ‘feel’ than local analysis
      - Response time
      - Batch-like access capabilities (submit, cancel, monitor, track,…)
      - Single but flexible way to express / write their “job files”
- How to allow
  - For code to be installed transparently “enough”
    - on an heterogeneous Grid
    - Install “services” as needed

Build a strong and reliable Grid suitable for user analysis

And on time for data deluge in 2009?

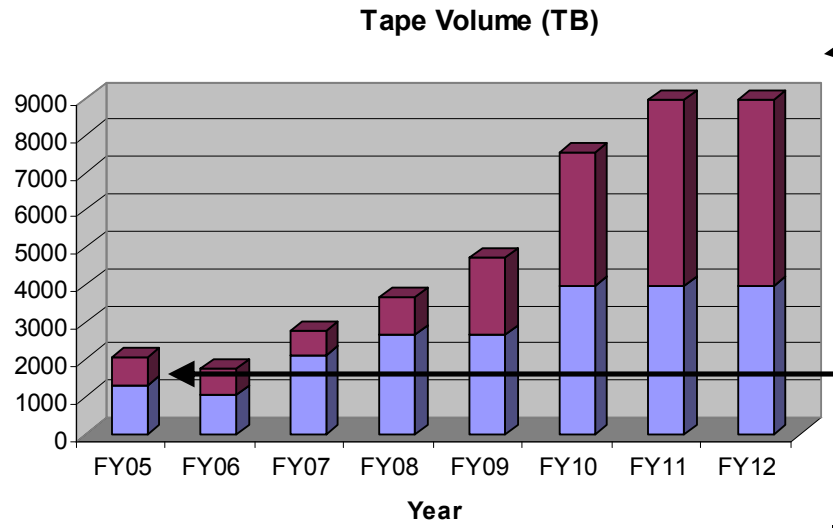
RHIC / RHIC II and all of that…

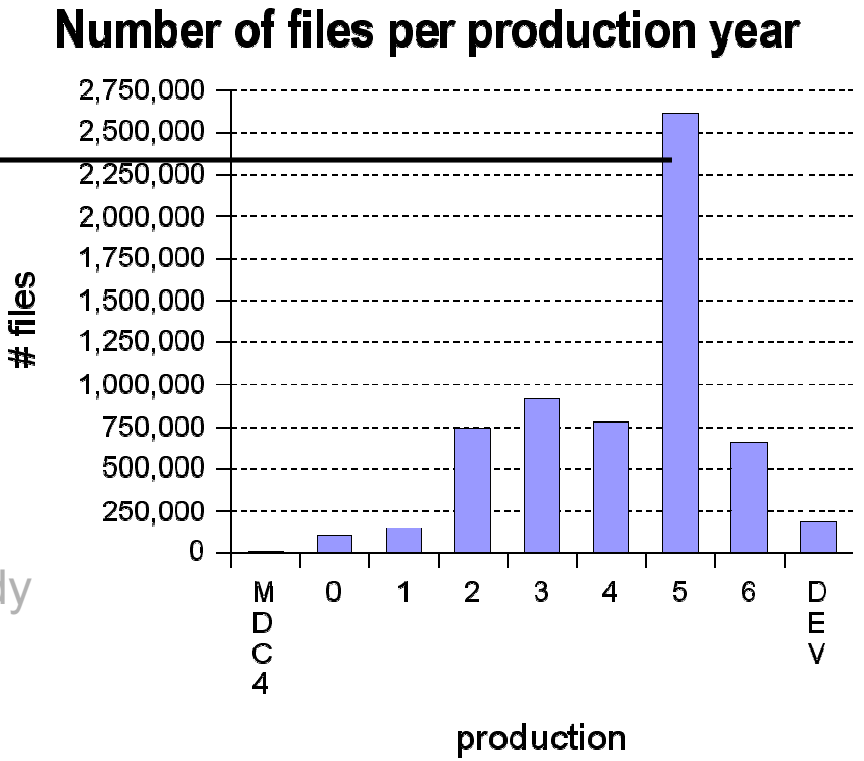
RHIC/ RHIC II and all of that…

- We have to fit within a global timing …
  - R&D and data deluge planned for 2009
  - STAR DAQ1000 project in place as prototype in THIS year experiment
    - Terrifically on schedule
  - Code development to achieve ‘from reading to tracking’ is on target
  - Grid goals should follow …
- Magnitude?

Scale of the problem

We need to be prepared for the DAQ 1000 era
One more order of magnitude





Number of files is set by RAW data sets and OS maximum file size
1 to 1 correspondence for Physics-ready files (ease data management)

Solutions?

- Develop strong Cataloging tools (operational basic)
- Develop strategy to distribute data immediately
  - Need for protocols and common interface
  - Implementation is a detail BUT common strategy
- Develop front end tightening data / computational needs
  - Independent of resource / hardware / platform choice
  - Local or distributed resources aware, resource brokering and planning
  - Knowledge of policies, limitations, …
  - Workflow
- Data access model
  - Reference abstraction
  - Local, remote, optimized “pool to pool” transfer
  - Advanced reservation, planning & prediction (shared network)
  - Quota, accounting
  - Object access model

So? Past, ongoing, future …

- Production use of SRM
  - DataMover + RRS (Replica Registration Service)
  - SRM file transfer on Grid (need tweak as new requirements imposed by CS are coming)
- Standard interface to
  - Job submission & control
  - Access to job (workflow) Monitoring information
  - Access to job (workflow) tracking information
- User analysis tools
  - Integrating of SRM technologies, corner stone to later plans
  - Xrootd, GridCollector all rely on SRM long life
  - OOD SAP aimed to allow Object level access (large dataset implies efficient/fast access to sub-sets on demand)
- Summary – In principle, the technology pieces exists – We want to
  - Leverage SRM technology
  - Integrate into other success story tools and provide the next generation of user ANALYSIS tools

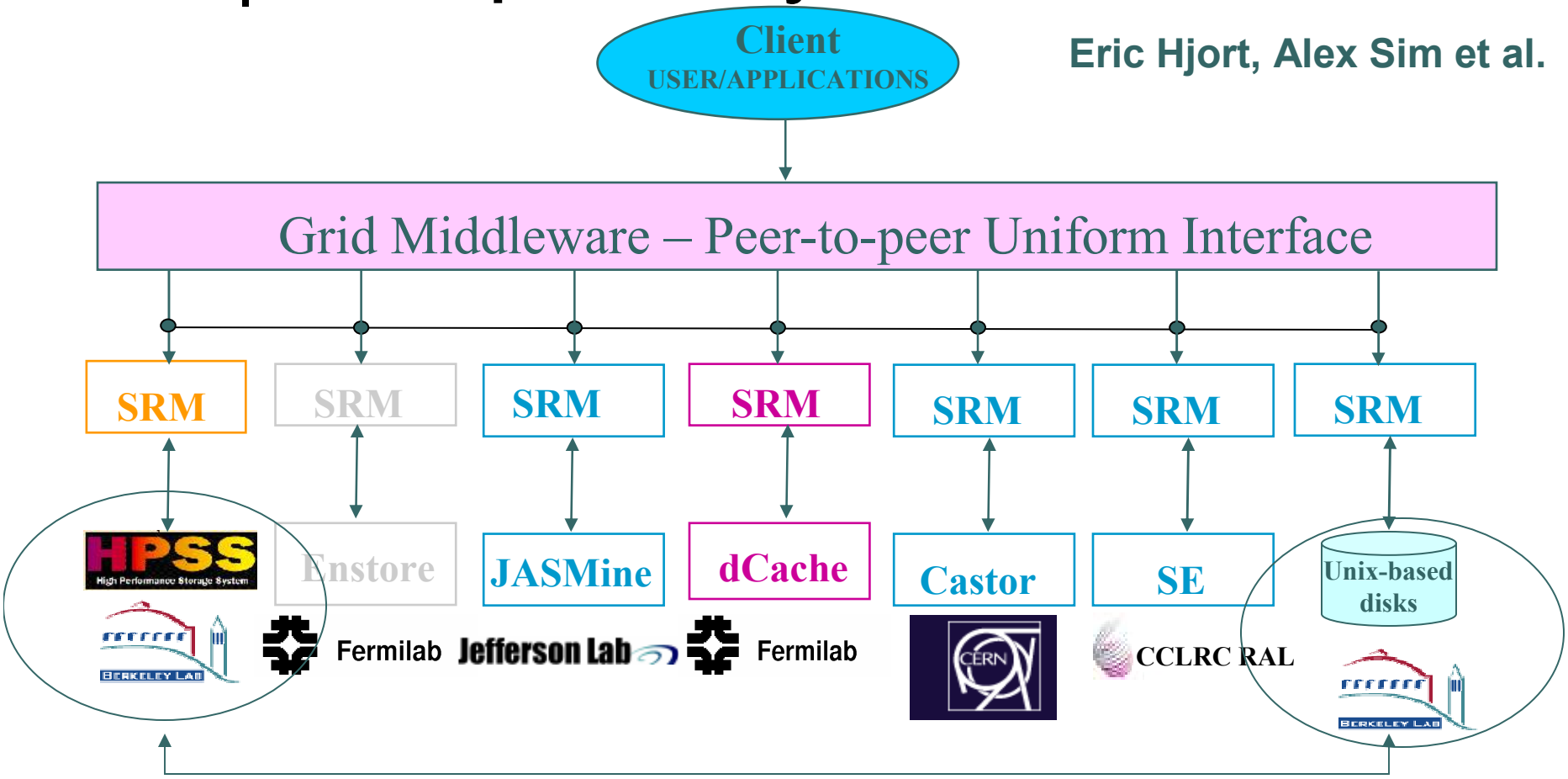
The products then and now

SRM, SRM and … SRM

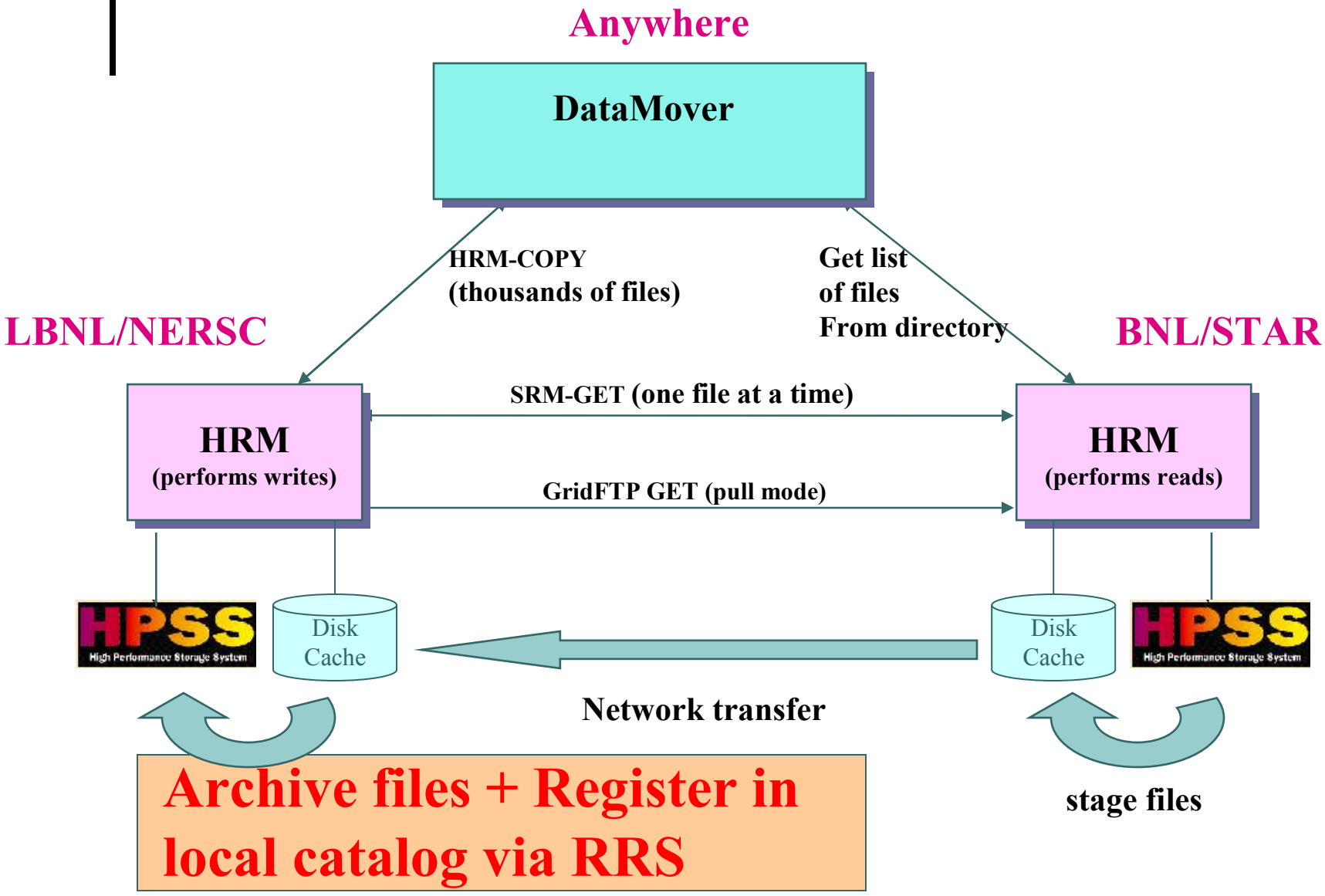
- SciDAC-1 told us – SRM is the way
  - We believed it
  - We benefited from it
  - We integrated it into our strategy
    - Assuming additional support for evolution, …
- SRM
  - Dynamic storage management
  - Only (tool) to provide dynamic space reservation
  - Resolve storage clogging issues
  - De-facto world wide standard
  - US SciDAC-1 success
  - Used in production by many experiments
- Current support level
  - **None direct**
  - **Grass-root only – how long is this sustainable?**
- Statements from STAR
  - **We ARE Data-Grid driven**
  - **We NEED data transfer tools with low failure rate and superior capabilities – SRM and the SRM collaboration provid{es|ed} it**

Uniformity of Interface ➔ Compatibility of SRMs

Eric Hjort, Alex Sim et al.



DataMover: HRMs use in PPDG- STAR (and ESG) for Robust Muti-file replication



SRM/DataMover Achievements

Eric Hjort, Alex Sim et al.

- Integrating Datagrid Technology with Physics Experiment End-to-end Applications.
  - PPDG News Update – 25 September 2002
  - 1 TB /week data transfer, planned 3 TB/week 2003
- Physics results from the STAR experiment at RHIC benefit from production Grid data services.
  - PPDG News Update – 19 Mar 2004
  - 5 TB/week production mode with Catalog registration data transfer coast to coast
  - Discrepancy rate 0.02% - 50 times less than before Grid solution
  - **Direct Quark Matter 04 impact, ½ of the analysis were done at NERSC/PDSF**
- Data transfers to NERSC/PDSF have ever since enabled an explosion of analysis by bringing data to the Physicists
  - **Recently, NERSC/PDSF a hub for data reduction & re-distribution to China**
  - **Direct Quark Matter 06 impact …**

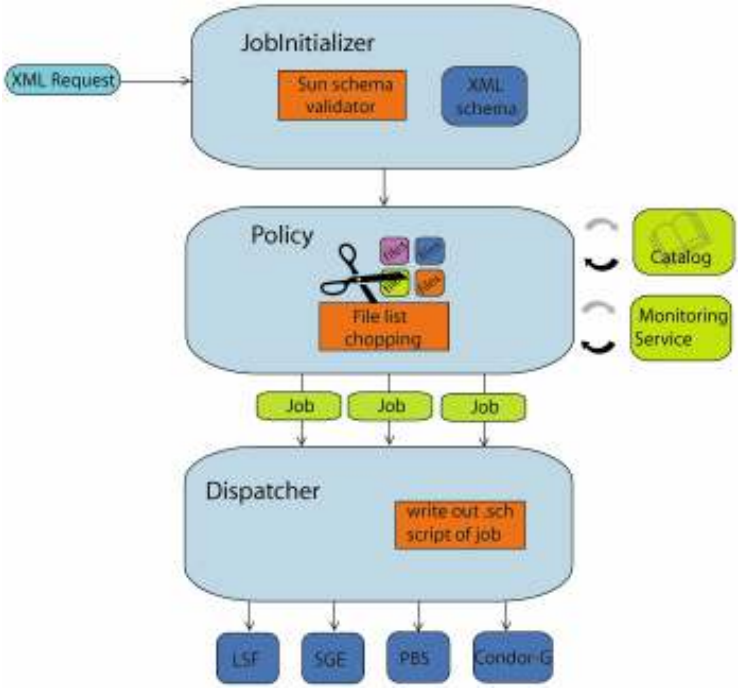
pion, kaon, proton v2 and v4 from ToF at 62.4 and 200: Xin Dong (**USTC**)
K0s and Lambda v2 and v4 at 62.4 and 200: Paul Sorensen (BNL), Yan Lu (**IOPP/LBL**)
K0s and Lambda Lee Yang Zeroes high pt v2: Yan Lu (**IOPP/LBL**)
High pT Pion and Proton v2 Using rrdEdx at 62.4 and 200: Paul Sorensen (BNL)
Xi and Omega v2 at 200: Kai Schweda (Heidelberg/LBL)
Xi and Omega v2, v4, and Centrality Dependence at 62.4 and 200: Markus Oldenburg (CERN/LBL)
phi v2 at 62.4 and 200: Sarah Blyth (U. of Capetown/LBL)
Kstar v2: Xin Dong (**USTC**)
Kstar Spin Alignment: Zebo Tang (**USTC**)
phi spin alignment: Jinhui Chen (UCLA)
Elliptic Flow Fluctuations: Paul Sorensen (BNL)
Identified Particle Correlation Studies: Jiaxu Zuo (**BNL/SINAP**)
Non-photonic Electron Flow Analysis: Andrew Rose (LBL)
CuCu v2: Nathan Beckett (LBL/Columbia)

SUMS

The STAR Unified Meta-Scheduler

Gabriele Carcassi
Levente Hajdu
…

- **STAR Unified Meta-Scheduler**
  - Lots of Architecture work to make it flexible
    - Tool is adaptable to any catalog, scheduler, policies, …
    - Plug-an-play architecture
  - **Meta = handshake with**
    - **ANY scheduler** (batch systems), local or distributed
    - **ANY analysis** (user, production, local or distributed, CPU intensive or IO intensive)
  - Gateway to user batch-mode analysis – Grid AWARE
- **Has allowed to optimize resource usage**
  - Not CPU only, Access to distributed 130 TB of distributed disk comparing to ~ 70 TB central (NFS, PANFS)



SUMS Achievements

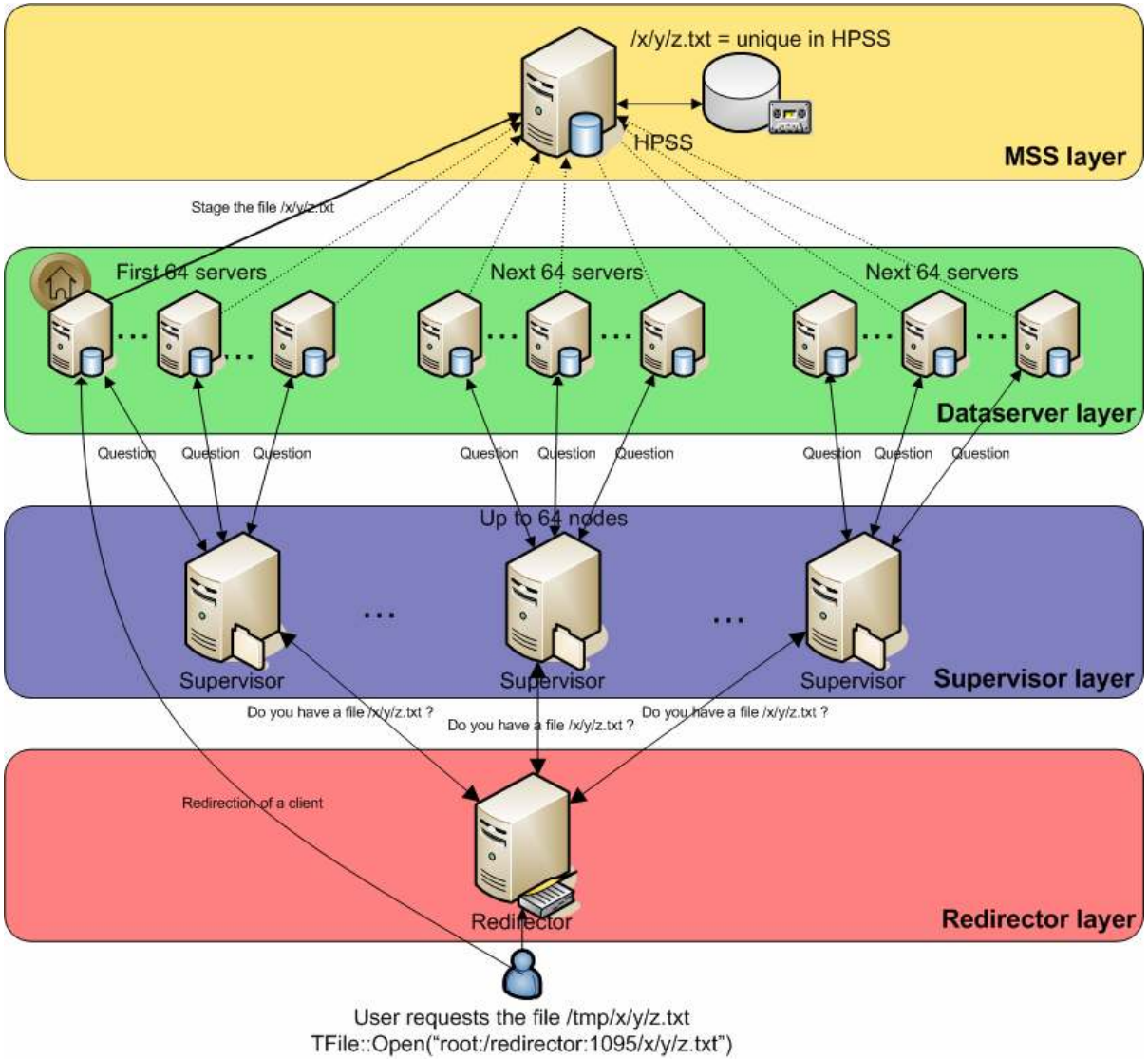
- Difficult to quantify, nonetheless ..
  - "Simple" - All data analysis done through SUMS
    - Technology changes DO not affect/delay Physicists
      - Migration through two batch systems at PDSF
      - Kick start analysis at WSU, SPU, BHAM, UIC (no change from user stand point)
      - Same interface for ANY local jobs
  - Subjective
    - Data sets have grown by one order of magnitude
    - Data sets are "hidden" on storage local to compute nodes
    - Analysis are more complex but analysis productivity is equal to greater than before
- Less trivial
  - Virtualization is within "understanding" – **U-JDL / RDL**
  - **Outreach**: Other Grid access
    - XGrid experience at MIT
      Linux centric, STAR can now support MAC-OS AND (the non trivial part), vendor Distributed Computing software stacks
    - Sun-Grid experience
      - We are running simulations on a commercial Grid, and morphing SUMS to map to its new coming interface

Data access through Xrootd

Client / server arch at
Its core

Re-director achieve
delegation of IO

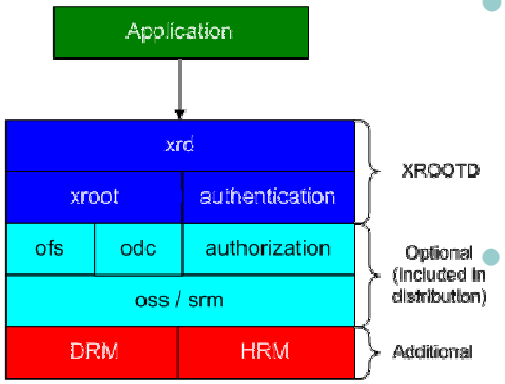
“Tree” of request
builds up, forming a
grid of data aggregation.



Xrootd choice & plans

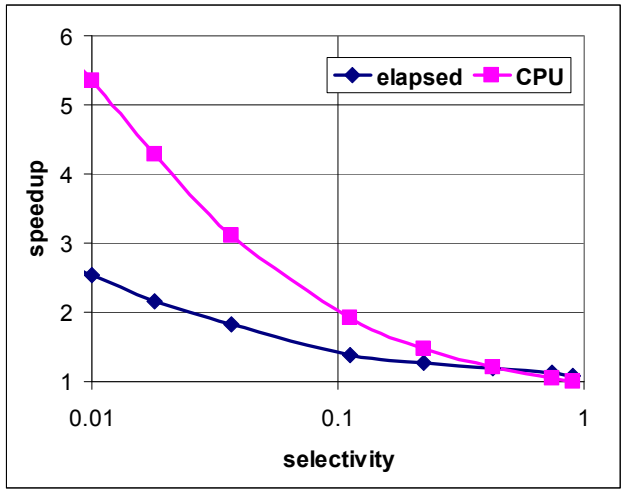
- STAR data mining, simulation, analysis IS based on ROOT framework
- STAR started with “rootd” (2003), moved to Xrootd (2005)
  - A highly scalable, self-configurable, fault-tolerant, plug-and-play component architecture tool suitable for technology evolution with ability to move hand-shake with Mass Storage Systems
  - **Cost effectiveness**
    - Low human maintenance cost (< 1 FTE)
    - Used in STRA to aggregate storage on computing nodes Hardware: order of magnitude (5-10) cheaper than leading centrally available solution
  - Impact
    - 64 TB of centralized disk , 134 TB on distributed , data mostly on distributed disk IO aggregate scales linearly with data-servers
    - Exceeds industry leading NAS&SAN i.e. analysis aggregate have faster turn around (TBC)
- Our plan
  - XROOTD+SRM
  - Second prototype is in test mode
  - Integration in STAR environment by February

From rootd to Xrootd, from physical to logical files: experience on accessing and managing distributed data
P. Jakl et al. (CHEP06)



GridCollector
Speeding up analysis

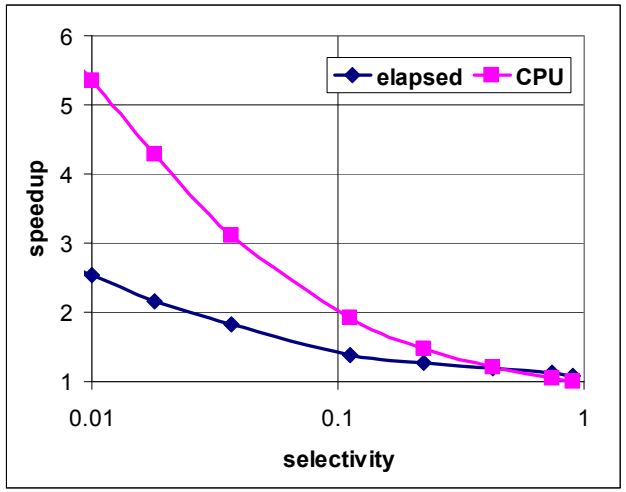
- Rests on now well tested, deployed and robust SRM
  - **Next generation of SRM based tools,** files moved via SRM service
  - Immediate Access and managed storage space (*single cache*)
- Easier to maintain, prospects are enormous
  - "Smart" IO-related improvements and home-made formats no faster than using GridCollector (a priori)
    - Hidden implication: **we can work long and hard and will unlikely do better**
    - Physicists could get back to physics



- Legend
  - Selectivity: fraction of events needed by the analysis
  - Speedup = ratio of time to read events without GC and with GC
  - Speedup = 1: speed of the existing system (without GC)
- Results
  - When searching for rare events, say, selecting one event out of 100 (selectivity = 0.01), using GC is 2.5 to 5 times faster
  - One order of magnitude more selectivity showed speed-ups by 20 to 50
  - Even using GC to read 1/2 of events, speedup > 1

GridCollector
Speeding up analysis

Kesheng Wu, Junmin Gu, Jerome Lauret, Arthur M. Poskanzer, Arie Shoshani, Alexander Sim, and Wei-Ming Zhang, Grid Collector: Facilitating Efficient Selective Access from Data Grids. In Proceedings of International Supercomputer Conference 2005, Heidelberg, Germany. *Best Paper Award*



- Legend
  - Selectivity: fraction of events needed by the analysis
  - Speedup = ratio of time to read events without GC and with GC
  - Speedup = 1: speed of the existing system (without GC)
- Results
  - When searching for rare events, say, selecting one event out of 100 (selectivity = 0.01), using GC is 2.5 to 5 times faster
  - One order of magnitude more selectivity showed speed-ups by 20 to 50
  - Even using GC to read 1/2 of events, speedup > 1

Still needed

- SRM and multi-users will soon need
  - Role-based and later user authorization, quota
  - Best data placement, data transfer queue monitoring
  - Policies, …
- Much needs to be done for data analysis
  - Next scale data challenge (DAQ1000 era) + Physics of Rare probes
    - Will require efficient data selector like GridCollector
  - Emerging complementary techniques – need access to “quanta” (events? Events?)
    - Xrootd provides redirection, scalable IO, aggregation
    - GridColletcor provides object access methodology
    - Xrootd, Proof, GridCollector may benefit from merging – Xrootd+SRM on the way
    - Calls for an **Object on Demand system**
  - Would be of a benefit for ANY community using the ROOT framework (most HEP/NP exp.)

Requirements?

- SRM
  - Require 0.02% or less failure on data transfer
  - Need proxy-ing through GK
  - Need enhanced features – policies, quota, user based authorization
- "Batch" job support
  - Need standard interfaces to submit jobs
    - Support for popular batch systems (SGE, …) and as they come
  - Would benefit from U-JDL / RDL integrated as a submission service including workflow
  - Need ways to monitor, track, control jobs on the Grid
    - Without CTRL-ALT-K, stand on one foot and pray
    - WITH standard interfaces and 21st century technology aligned with user's expectations
- Redistribution & packaging support
  - Should include environment and be solid based on OS version
  - Should be able to deploy services on GK (only node speaking in/our of a cluster)
- Operation
  - Need sustainable and scalable operation (STAR site expansion beyond expectation by x2)
- Infrastructure stability
  - >= 80% infrastructure efficiency for simulation
  - >= 85% infrastructure efficiency for data mining
  - >= 90% infrastructure efficiency for user analysis
  - Support for 10k jobs a day
  - Obviously need error recovery
- Infrastructure flexibility
  - Need information about storage and need to be able to "grab" it
  - Need to be able to install persistent service (Xrootd, SRM, …) open ports for protocol support
- Infrastructure scalability
  - Require flexible "resource" pining coupled with job throttling
- *+ Would require scalable database (clearly dev)*
  - *Self discovering, distributed database*
  - *15 M records search space within 5 seconds or less*
  - *100 M within 10 seconds or less*
- *- Sent space requirements ages ago …*

Sites

Site	CPU Slots	Disk Space (GB)	GK name
NERSC-PDSF	1028	12298	pdsgrid2.nersc.gov
STAR-BNL	678	17211	stargrid02.rcf.bnl.gov
UIC-PHYSICS	96	190	mstr1.cluster.phy.uic.edu
STAR-BHAM	92	1398	rhilxs.ph.bham.ac.uk
STAR-SPU	14	56	stars.if.usp.br
STAR-WSU	7	229	rhic23.physics.wayne.edu

- BHAM, UIC x2
- SPU + 100
- One more site soon in South America
- Conact in China (interested)

OSG Milestones & extensions

STAR: Migration of >80% of simulation to OSG

- **F.1 Phase I: Months 1-18**
  - STAR: Migration of all (most) simulation to an OSG based operation, use of opportunistic resources with a combined software packaging and *deployment and on-the-fly SRM deployment*.
- **F.2 Phase II: Months 19-36**
  - STAR: Support for user batch analysis on the distributed facility, object based and interactive analysis for the STAR collaboration. Scaling of the order of 10k jobs/day and beyond as well as a robust infrastructure will need to be reached by the second 1/3rd of Phase II.
  - Extensions
    - Deploy a high-level user language to consistently express user interactive analysis and batch-based workflow.
- **F.3 Phase III and IV: Months 37-48 and 49-60**
  - Create virtual-machine-based sandbox for safe testing on an extended Grid and easy install and teardown of lab resources.