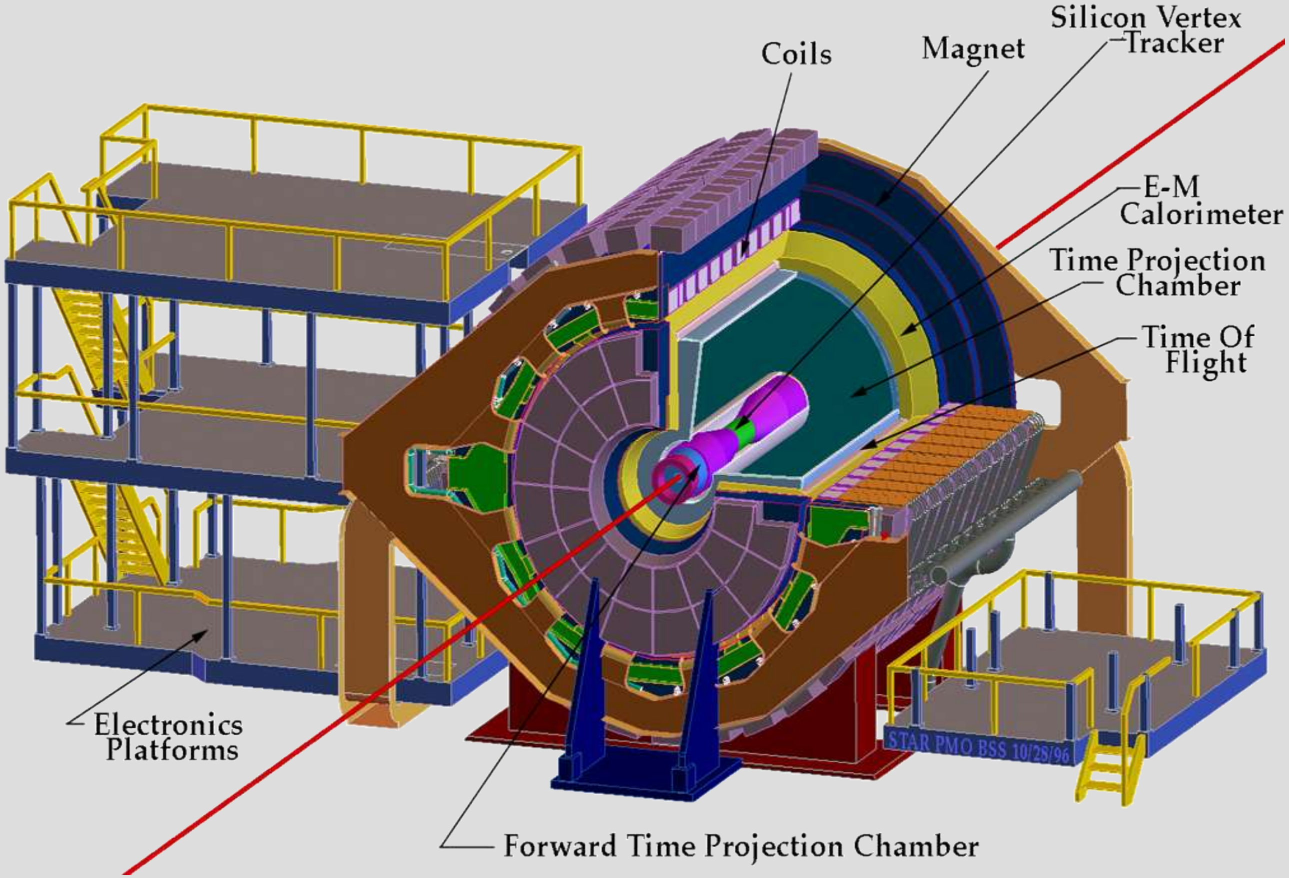
The RHIC/STAR Experiment and the SDM center

Jérôme Lauret
jlauret@bnl.gov



Outline

- ✓ RHIC experiments, scientific program, data and scales
- ✓ The past needs
- ✓ SDM projects & use in STAR & impact
- ✓ Global impact on the community
- ✓ The Future
- ✓ Thoughts

Outline

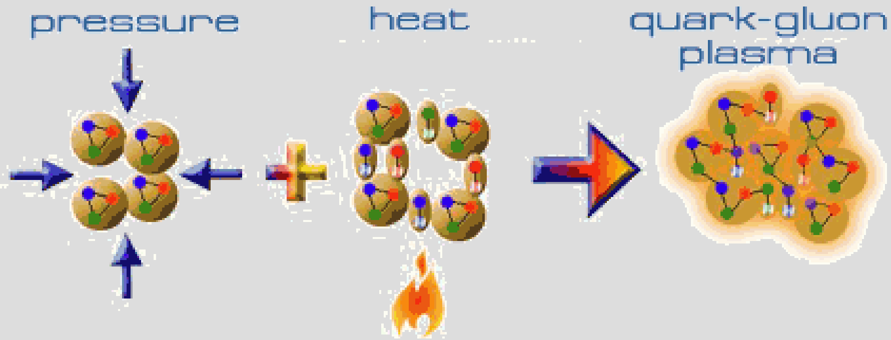
- RHIC experiments, scientific program, data and scales
- The past needs
- SDM projects & use in STAR & impact
- Global impact on the community
- The Future
- Thoughts

The Experiment(s) / programs

Sponsored by the U.S. Department of Energy
Office of Science - Office of Nuclear Physics.

- **RHIC = Relativistic Heavy Ion Collider**
  - An experiment located at the Brookhaven National Laboratory
  - study what the universe may have looked like in the first few moments after its creation
  - Current RHIC experiments: **STAR**, PHENIX, BRAHMS, PHOBOS (http://www.bnl.gov/RHIC/)

- **What do we do ??**
  - Heavy Ion smashing machine ...

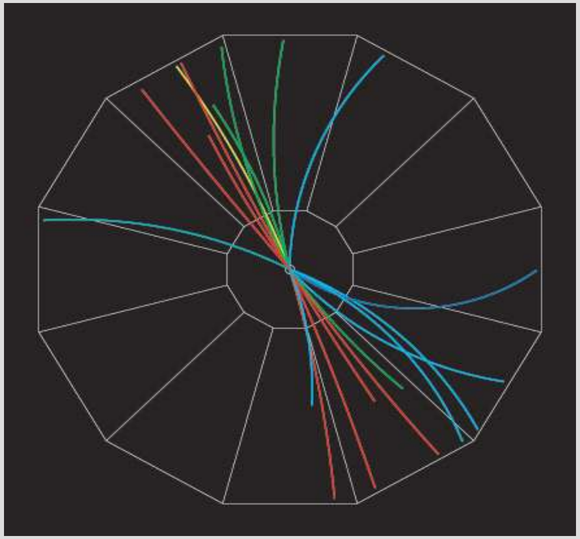


The Experiment(s) / programs

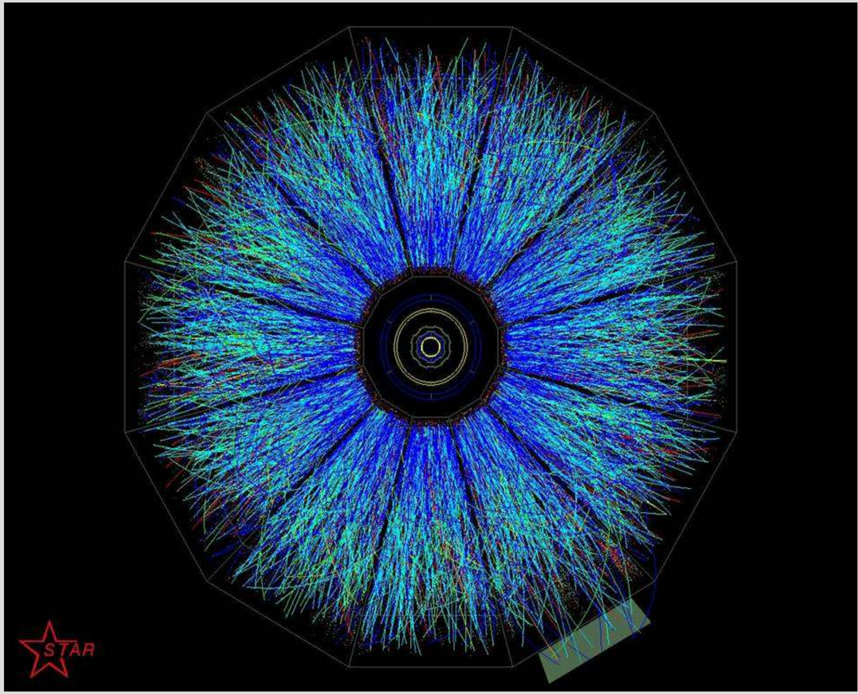
Sponsored by the U.S. Department of Energy
Office of Science - Office of Nuclear Physics.

- **What do we study ??**
  - The properties of the QCD under extreme conditions: (de) confinement, gluon saturation, phase transition, new State of matter, jet quenching, elliptic flow, partonic energy loss, ... spin structure of nucleons
  - Rare probes, rare signals
  - In events of 1-2 MB, Million scale possibly going to billion scale starting from 2008

We need to find this

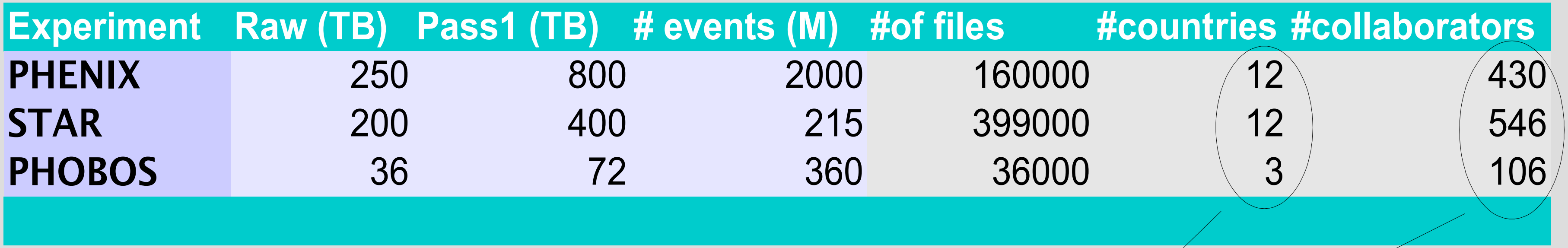


in this ... knowing that it comes once every N events.



Setting the scale
Our data

2003/2004 data



Experiment	Raw (TB)	Pass1 (TB)	# events (M)	#of files	#countries	#collaborators
PHENIX	250	800	2000	160000	12	430
STAR	200	400	215	399000	12	546
PHOBOS	36	72	360	36000	3	106

General observations

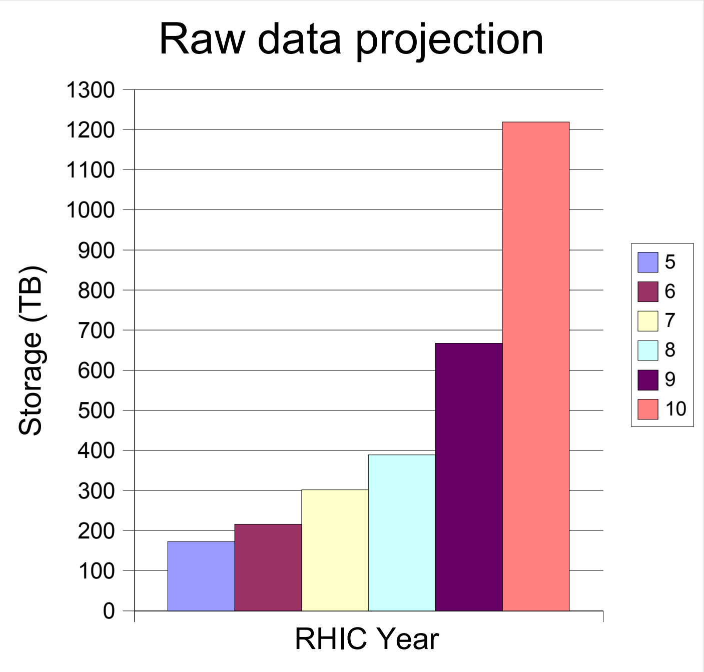
- Many countries, continent, collaborators
- PB scale (overall and including reconstruction pass)
- Large amount of files (overall reaches millions)

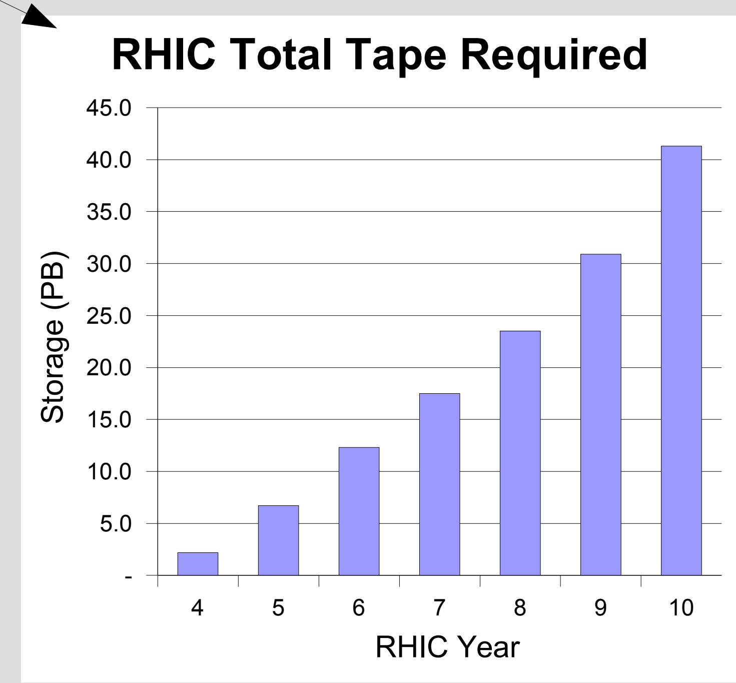
Setting the scale
Our data

2003/2004 data

Experiment	Raw (TB)	Pass1 (TB)	# events (M)	#of files	#countries	#collaborators
PHENIX	250	800	2000	160000	12	430
STAR	200	400	215	399000	12	546
PHOBOS	36	72	360	36000	3	106

Projections





A few field / research realities
A starting point

- **File based analysis**
  - Started with more resources than necessary
  - Files contained pre-triggered (or filtered) events
  - A file is always part of a larger “collection”
  - A collection is defined by
    - The run configuration, a set of triggers used in the run, the sub-detectors present, ...

- **Analysis statistically driven for the most part**
  - BE AWARE of time sorted event sequence for some analysis
    - Usually missed in most CS projects, time dependent fluctuation implies subtle event correlations
    - Has DRASTIC design consequence for event-based servers
- **ROOT based frameworks**
  - Currently common to RHIC experiments (http://root.cern.ch/)

Outline

- RHIC experiments, scientific program, data and scales
- The past needs
- SDM projects & use in STAR & impact
- Global impact on the community
- The Future
- Thoughts

Experiment needs then

Driven by

- Large amount of projected data & files (now a reality)
- Collaborations spanning over 12 countries, distributed resources, increasing demand

Early (or semi-early depending on strategy)

- Cataloging of files
  - There was hope for a general Replica/File/MetaData-Catalog from the Grid landscape but it did not come.
  - Experiments developed their own

Experiment needs then

Driven by

- Large amount of projected data & files (now a reality)
- Collaborations spanning over 12 countries, distributed resources, increasing demand

Early (or semi-early depending on strategy)

- Cataloging of files
  - There was hope for a general Replica/File/MetaData-Catalog from the Grid landscape but it did not come.
  - Experiments developed their own
- File transfer strategy →
  - Only happened for experiments having from the start +1 site
  - SRM / DRM / HRM

Experiment needs then

Driven by

- Large amount of projected data & files (now a reality)
- Collaborations spanning over 12 countries, distributed resources, increasing demand

Early (or semi-early depending on strategy)

- Cataloging of files
  - There was hope for a general Replica/File/MetaData-Catalog from the Grid landscape but it did not come.
  - Experiments developed their own
- File transfer strategy
  - Only happened for experiments having from the start +1 site
- Cataloging of events → SDM
  - Extremely rare use in our field (a shame really)
  - Bitmap index, STACS, GridCollector

Experiment needs then

Driven by

- Large amount of projected data & files (now a reality)
- Collaborations spanning over 12 countries, distributed resources, increasing demand

Early (or semi-early depending on strategy)

- Cataloging of files
  - There was hope for a general Replica/File/MetaData-Catalog from the Grid landscape but it did not come.
  - Experiments developed their own
- File transfer strategy
  - Only happened for experiments having from the start +1 site
- Cataloging of events
  - Extremely rare use in our field (a shame really)
- Access to distributed computing and/or storage resources
  - Discovery & Efficient access of resources
  - Scheduler, planner, ...

Outline

- RHIC experiments, scientific program, data and scales

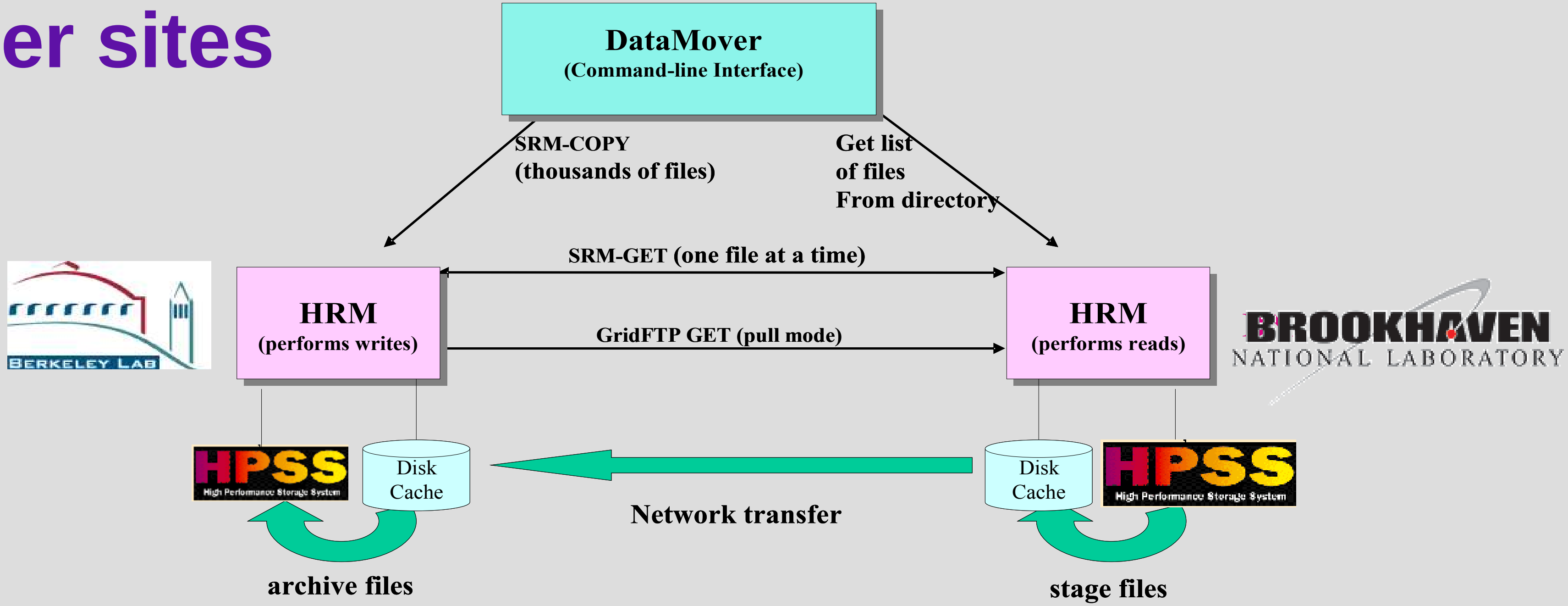
- The past needs

- SDM projects & use in STAR & impact

- Global impact on the community
- The Future
- Thoughts

Data transfer in STAR
SDM Data-Mover

- **STAR started with**
  - A Tier-0 site - all “raw” files are transformed into pass1 (DST), pass2 (MuDST) files
  - Tier-1 site - Receives all pass2 files, some “raw” and some pass1 files
- **STAR is moving toward replicating this to other sites**



DataMover
(Command-line Interface)

SRM-COPY
(thousands of files)

Get list
of files
From directory

SRM-GET (one file at a time)

GridFTP GET (pull mode)

HRM
(performs writes)

HRM
(performs reads)

BERKELEY LAB
BROOKHAVEN
NATIONAL LABORATORY
HPSS
High Performance Storage System
HPSS
High Performance Storage System

Disk
Cache

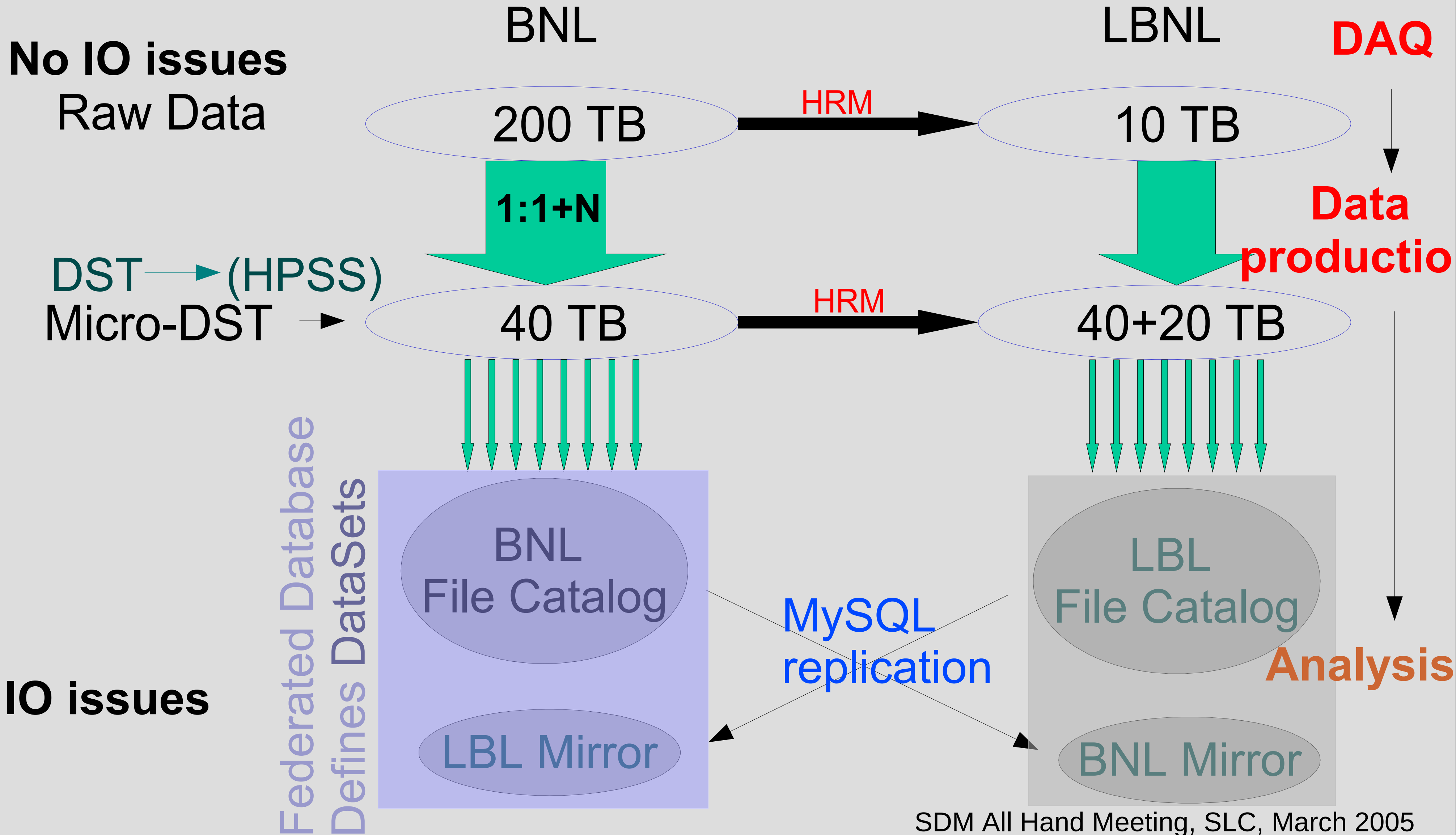
Disk
Cache

Network transfer

archive files

stage files

Data transfer flow



No IO issues

Raw Data

BNL

LBNL

200 TB → HRM → 10 TB

1:1+N

DST → (HPSS)

Micro-DST → 40 TB → HRM → 40+20 TB

Federated Database

Defines DataSets

BNL File Catalog

LBL Mirror

MySQL replication

LBL File Catalog

BNL Mirror

IO issues

DAQ

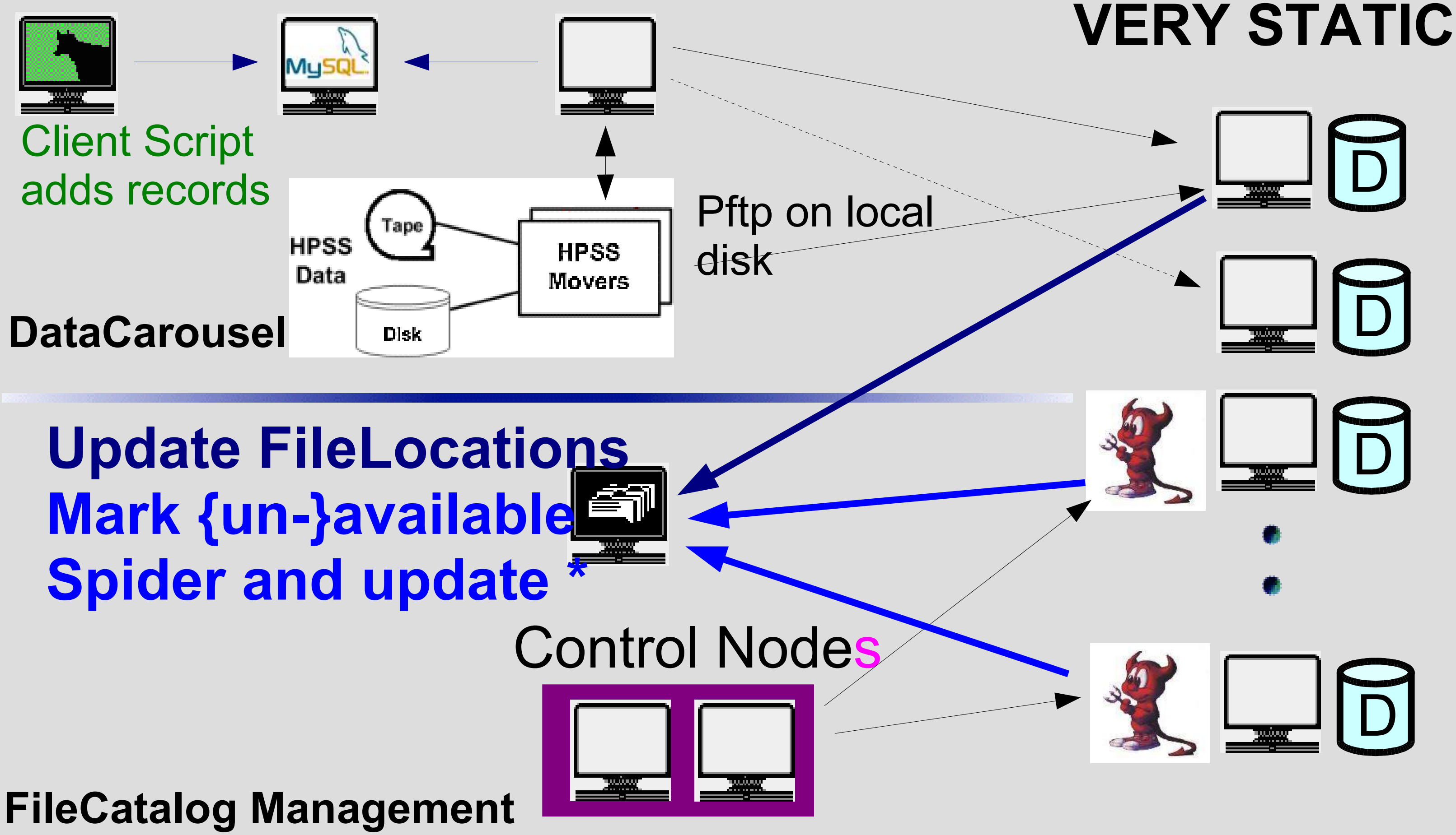
Data productio

Analysis

Data transfer flow

Where does this data go ??

VERY HOMEMADE
VERY STATIC



Client Script adds records

DataCarousel

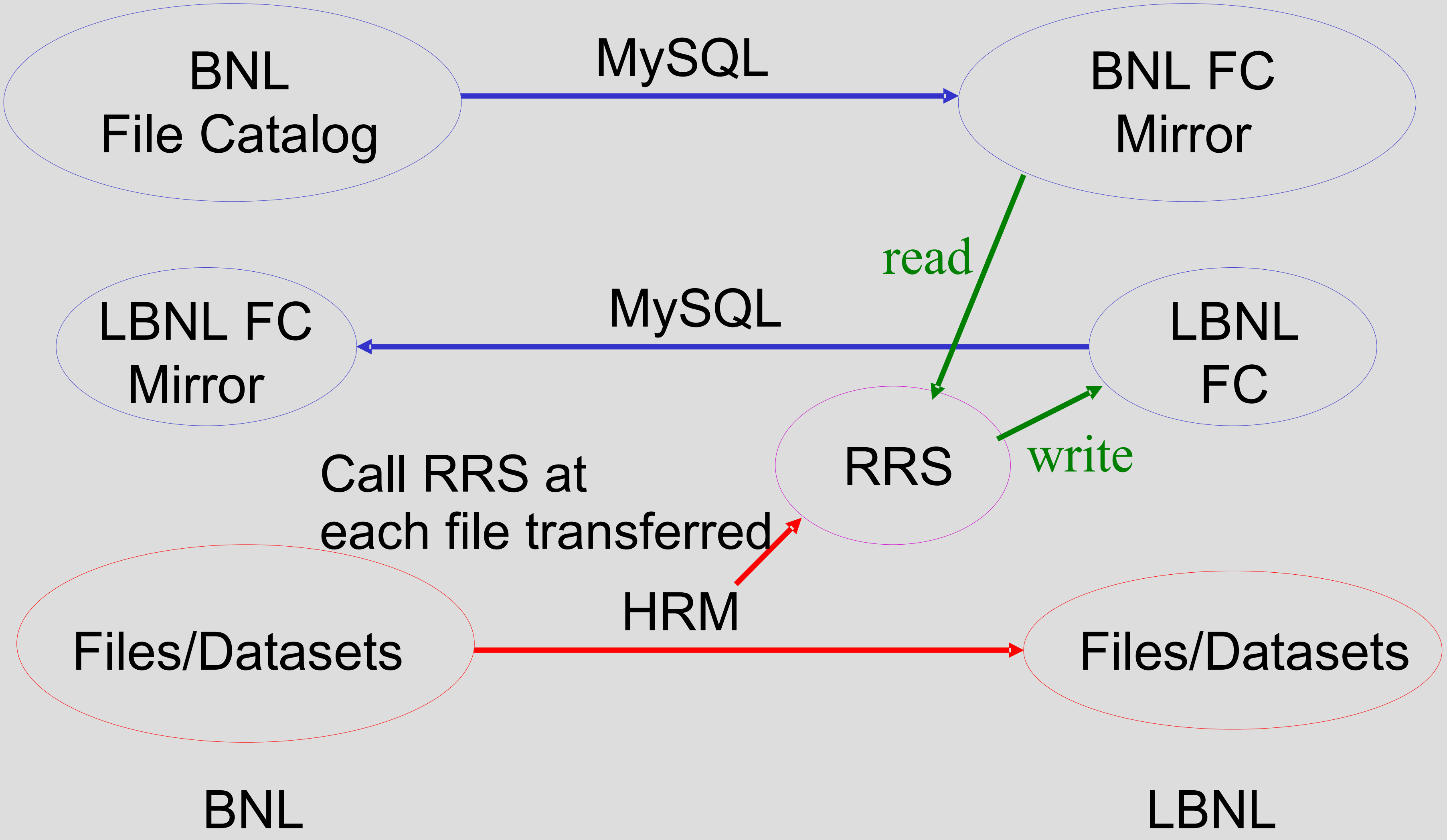
Pftp on local disk

Update FileLocations
Mark {un-}available
Spider and update *

Control Nodes

FileCatalog Management

Data transfer flow



Experience with SRM/HRM/RRS

- **Extremely reliable**
  - Ronko's rotisserie feature “*Set it, and forget it !*”
  - Several 10k files transferred, multiple TB for days, no losses
  - Project was (IS) extremely useful, production usage in STAR
  - Data availability at remote site as it is produced
    - We need this NOW (resource constrained => distributed analysis and best use of both sites)
    - Faster analysis yield to better science sooner
    - Data safety
- **Since RRS (prototype in use ~ 1 year)**
  - 250k files, 25 TB transferred AND Cataloged
  - 100% reliability
  - Project deliverables on-time

In our book, it qualifies as success

GridCollector

CHEP04 "*Using an Event Catalog to Speed up User Analysis in Distributed Environment*"

- "tags" (index) based, need to be define a-priori [production]

- **Historically**
  - Its first incarnation: STACS, Grand-Challenge
    Did not really take-off ... Many reasons for that
    - There was no need (no resource constraints)
      - Project came too early
    - There were some functionality issues (not easy to rebuild the index, had to restart from scratch)
    - Manpower for support was not preserved (slow interest)

All and behold, users did not want to make the effort to use it and had no needs either

GridCollector

- **Current situation**
  - For one thing, time has come ... resource ARE constrained
  - Rest on now well tested and robust SRM (DRM+HRM) deployed in STAR anyhow
  - Manpower was found, interest generated
    - Hand on contact Collaboration with Kensheng was positive
    - Suggests closer collaboration when projects are more abstract
  - Easier to maintain, prospects are enormous
    - “Smart” IO-related improvements and home-made formats no faster than using GridCollector (a priori)
      - Physicists could get back to physics
      - And STAR technical personnel better off supporting GC

It is the only working prototype of Grid analysis framework - This is under-sold

GridCollector

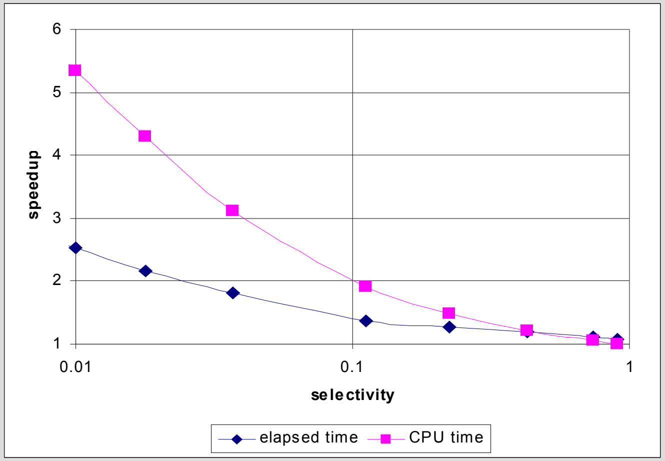


Figure 1: Using Grid Collector reduces both CPU time and elapsed time, and speeds up analysis jobs.

Not only gain of 40% but also

- Manages my space in a dynamic fashion
- Only done in a static fashion for now ... need dynamic since data availability > disk space

Outline

- ✓ RHIC experiments, scientific program, data and scales
- ✓ The past needs
- ✓ SDM projects & use in STAR & impact
- ✓ Global impact on the community
- ✓ The Future
- ✓ Thoughts

SRM/HRM/RRS

- **STAR benefits from SRM now**
  - SRM enabling sites allows for timely data arrival
  - Local resources usable immediately (was the goal)
  - More complex Tier0/Tier1/Tier2 layout is "*pre-tested*" with real life case scenario and practical deployment

- **SRM is becoming a de-facto standard**
  - OSG adopted, this need to be solidified and scaled up
  - SRM could be a generally accepted "service" (and the first generic Grid service too)
  - Immediate benefit to science
    - No need to re-invent or change our scheme(s)
    - Experiment Grid research could (should) move to new topics

The proof has been made, it has been done, need scalability testing

GridCollector

- **Impact are uncertain yet**
  - Landscape changes rapidly, user interest has to be maintained
  - Pre-analysis seem to demonstrate benefit
    - Will need publication to back this up

- **Nonetheless**
  - The time is the proper time
  - Second generation of SRM-based tools needed => GC
  - Event based analysis __IS__ the next frontier
  - Index bitmap generating interest amongst the wisest
  - Potential uses and consequence of having an event coordinator are endless

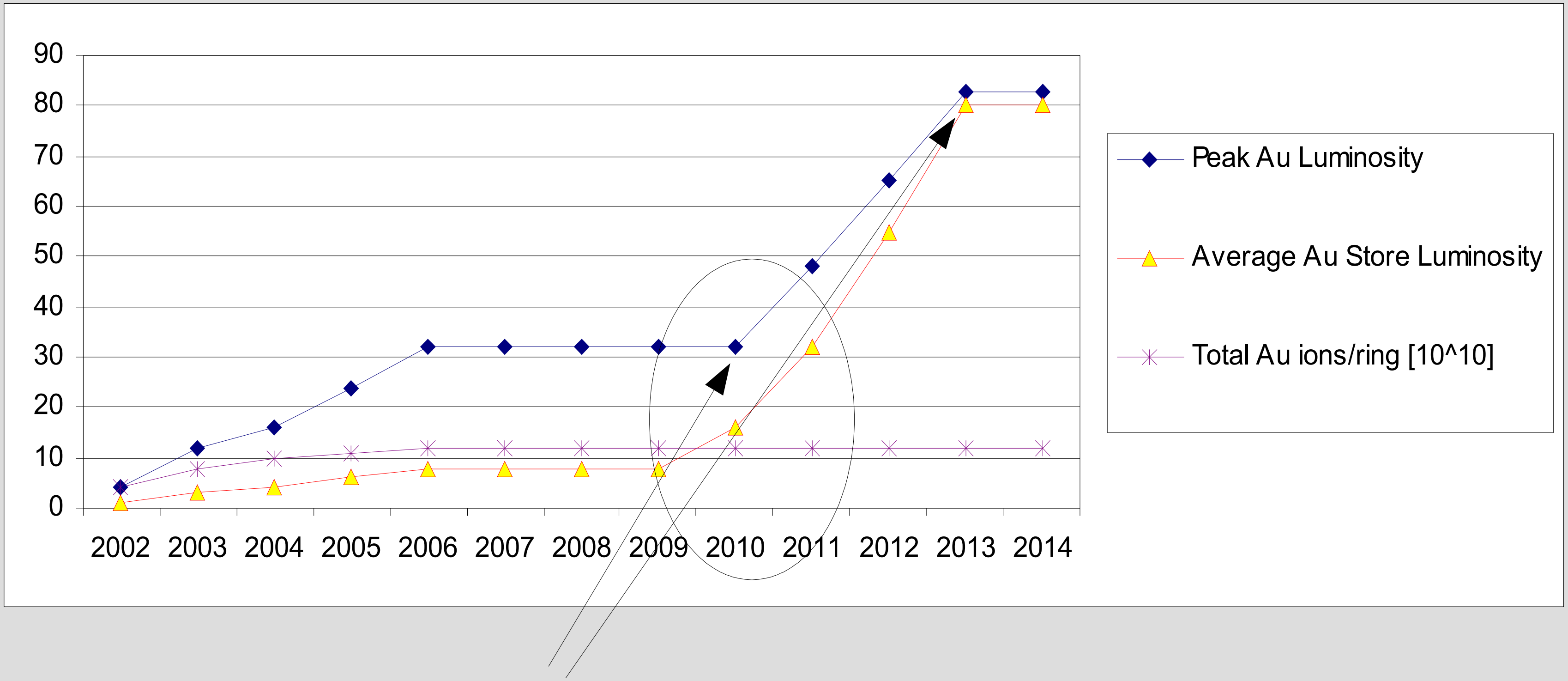
    Side story
  - Ongoing projects in STAR make me think “objectivity or GC ?” and this time, it is at event reconstruction level (raw format)

Outline

- ✓ RHIC experiments, scientific program, data and scales
- ✓ The past needs
- ✓ SDM projects & use in STAR & impact
- ✓ Global impact on the community
- ✓ The Future
- ✓ Thoughts

RHIC-II era

STAR planning for RHIC-II



The luminosity increase by X > 2 implies a possible data rate (and amount) of x2 or larger ...

In fact, the current few years are best suited for new projects and R&D to prepare for the later years ...

End of RHIC / RHIC-II era

- **RHIC-II does not mean x2 luminosity alone**
  - DAQ development (in STAR, x10 up to 1 GB/sec)
    - IO is also home made using streams o separate box – Room for efficient IO
  - Detector development
    - Some designed system are not synchronized with the rest of the event timeline and will need re-sequencing
    - Cataloging may be an issue
    - GC may make it easier to “bring the files we need” for reconstruction
  - Physics of rare probes
    - Intensified un-triggerable data sample
    - Requires up to billion of events
    - Resource availability make these prohibitive, GC opens the possibility again

End of RHIC / RHIC-II era

- **Expansion of Tier-1**
  - Continuous support for SRM based tools important (next generation)

- **Being able to use resources does not mean using them efficiently**
  - What does this mean for SRM, file placement ?? Mesh of SRM ?? “bring the file from the *best* site and *best* storage”
  - Interested in DLT for example (bandwidth at a cost drives scheduling and job split) & planner ideas
  - What about accounting, quotas, priorities in SRM ??
    - User are using DRM implicitly now through GC
    - Soon, it will be needed – Beware of the entropy

HENP time-line considerations

- **RHIC: next wave in 2010-2020+**
  - But plateau suitable for new development starting next year i.e. NOW. Shall we take advantage of it ??
  - We are moving to the second generation of SRM based tools ... there are immediate needs (previous slide) to bring confidence and next cycle.

- **LHC: start 2007 on a 20+ years program**
  - + 2/3 years LHC left before production mode
    - Priority need probably on current implementation scalability/stability
    - Interoperability required (day one implies 2 continents)
    - Assumed basic needs re RHIC-like needs
  - Alice VERY interested in bitmap index (analysis or vis.)
    - Do we work together on this ??
    - Will understand better soon ...

Outline

- ✓ RHIC experiments, scientific program, data and scales
- ✓ The past needs
- ✓ SDM projects & use in STAR & impact
- ✓ Global impact on the community
- ✓ The Future
- ✓ Thoughts

SRM/HRM/RRS ...

- **SRM becoming a de-facto standard ...**

  Pandora's box has been opened

  - May have its negative sides (??)
    - Development versus consolidation versus distribution
    - Maintainability and compliance from/to standards
    - Burden is higher
  - What is the state of Interoperability
    - dCache & SRM door (??)
    - Jlab SRM and LBNL SRM 100% inter-operable (??)
  - Need consolidation ?? How is this planned ??
  - Even if inter-operable, how to ensure continuity ??
    - RFC, protocol documentation in some official format, IANA service advertisement, ...
    - What if 50 experiments, 100 sites are using SRM ...

SRM, space aggregation ...

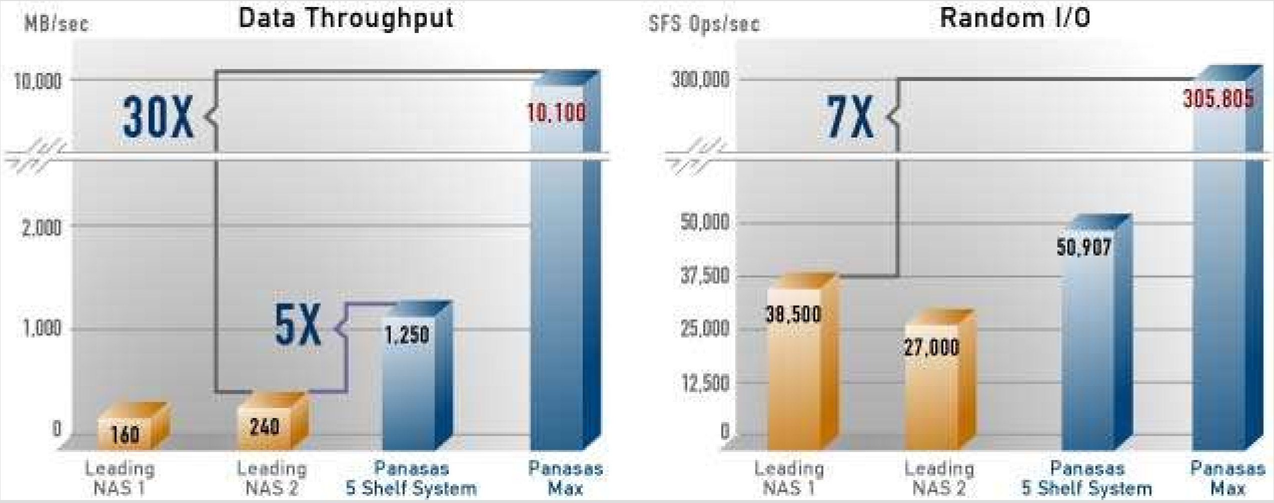
- **Different approach are emerging**
  - Xrootd ideas overlaps with SRM as per its space management ideas
    - Some components addresses aggregation of distributed storage
  - Same issue with global FS (Lustre, pvfs, ...)
    - How does this compete with professional solutions ??
  - Time to address it ?? Merger ??
  - Personal opinion
    - Worth a look (and the sooner the better)

Emergence of multiple coordinator makes access to HPSS (un-coordinated by nature) a disaster

Not sure from experiment stand-point which solution is best (just know that we need one)

IO a bottleneck but ....

- **PVFS interesting (aggregation, performance)**
  - Migration to from MSS a need
  - Transparent fast to slow storage migration would be nice
  - The question is to where (on the high side)



Finally, ...

- **What we could do better**
  - Project progress documentation could have been better
    - Non transparent / reproducible deployment
    - Need to work closer, STAR FTE / SDM + possibly doc and regular reports
  - Work often make big leaps after a trip, a common meeting or a visit on one side or another
  - Was hard to explain the DIRE need for RRS but finally got it (thanks!) - Balance of final solution / prototype ?

Conclusion (some)

- **Too many slides ;-) .. still a lot to do**
  - Still a lot to learn and knowledge to use from SDM
  - **Future is bright and allowing new development phase. If you want to work with us, now (or next year) is a good time**
- **Our preferences**
  - RMS question worth addressing ASAP (raising the flag like RRS)
  - Second generation to complete (GC, SUMS/SRM+RRS, more ...)
  - Consolidations
    - Scalability + Migration to stable technologies (Orbacus?)
    - Improvements (accounting, quotas, priorities, policies ...)
  - New development
    - Best placement, planner, scheduler issues, storage space aggregation
- **Don't know enough about**
  - Efficient IO, other data analysis, full use of bitmap index, ...
  - Here to learn too (so far was focused on data handling)