Real-time Data Access Monitoring in Distributed, Multi Petabyte Systems

Tofigh Azemoon

Jacek Becla

Andrew Hanushevsky

Massimiliano Turri

SLAC National Accelerator Laboratory

Soon a typical running HEP experiment will look like this

100s of users

↙ ↓ ↘

10s of thousands of batch nodes

↖ ↑ ↗

1000s of data servers

↑ ↑ ↑

10s of millions of files

Containing

10s of PB of data

+

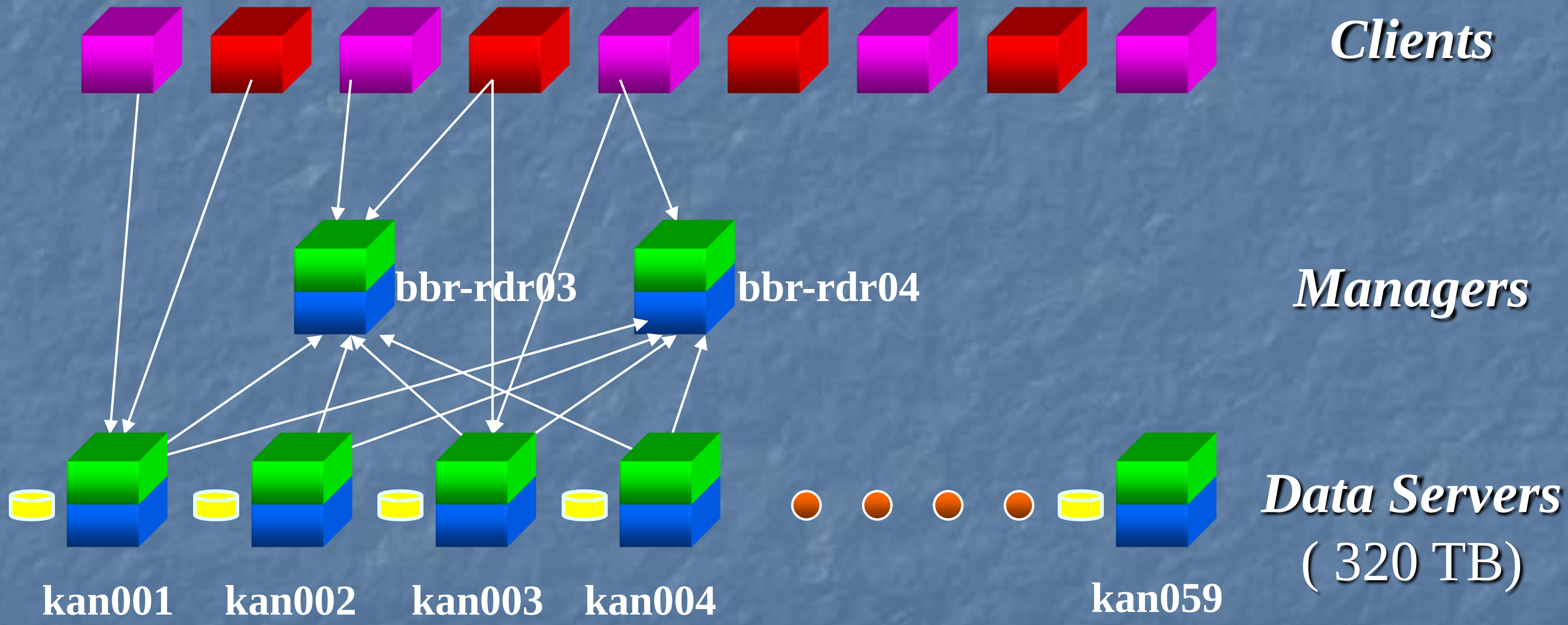
geographically distributed

crossing many time zones.

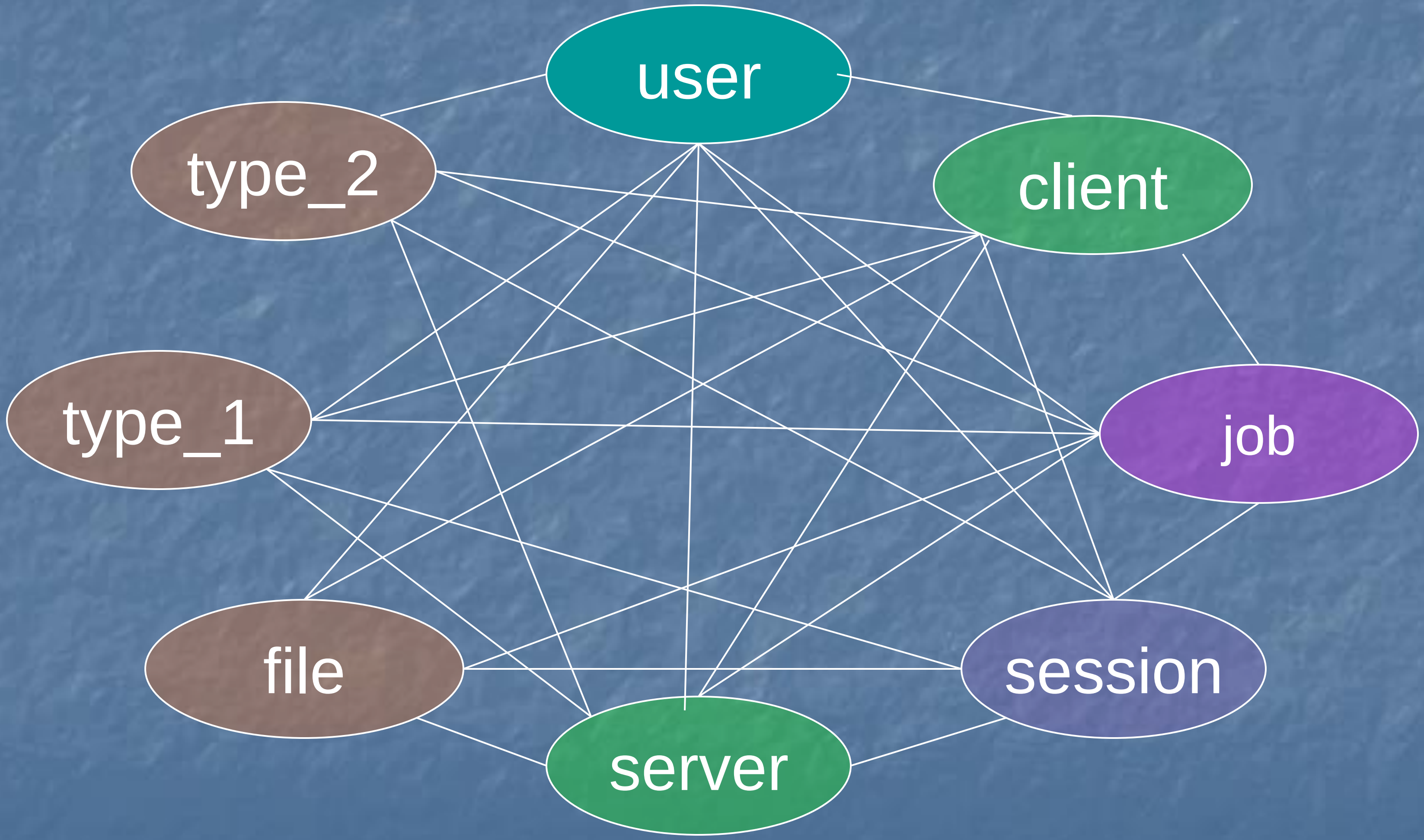
Mission Statement

- **Provide real time overall view of system performance**
- **Respond to detailed queries**
  - → **to identify bottle necks**
  - → **to optimize the system**
  - → **to aid in planning system expansion**

The SLAC ¼PB "kan" Cluster



Monitored Objects



File classes to monitor aggregate values for groups of files

BaBar Examples:

type_1 (dataType)

↓

/store/PR/R22/AllEvents/0006/70/22.0.3/AllEvents_00067045_22.0.3V03.02E.root

/store/SP/R22/000998/200406/22.0.3/SP_000998_068468.01.root

/store/PRskims/R22/22.1.1c/IsrIncExc/79/IsrIncExc_57978.01.root

/store/SPskims/R22/22.1.1c/Tau1N/001235/200212/Tau1N_001235_49553.01.root

↑

type_2 (skims)

File path → getFileType → (type_1 value,
type_2 value)

Xrootd Server

- Highly scalable server
- Posix like access to files
- Load balancing
- Transparent recovery from server crashes
- Fault tolerant
- Very low latency

Monitoring Implementation in xrootd

- Minimal impact on client requests
- Robustness in multimode failure
- Precision & specificity of collected data
- Real time scalability

→

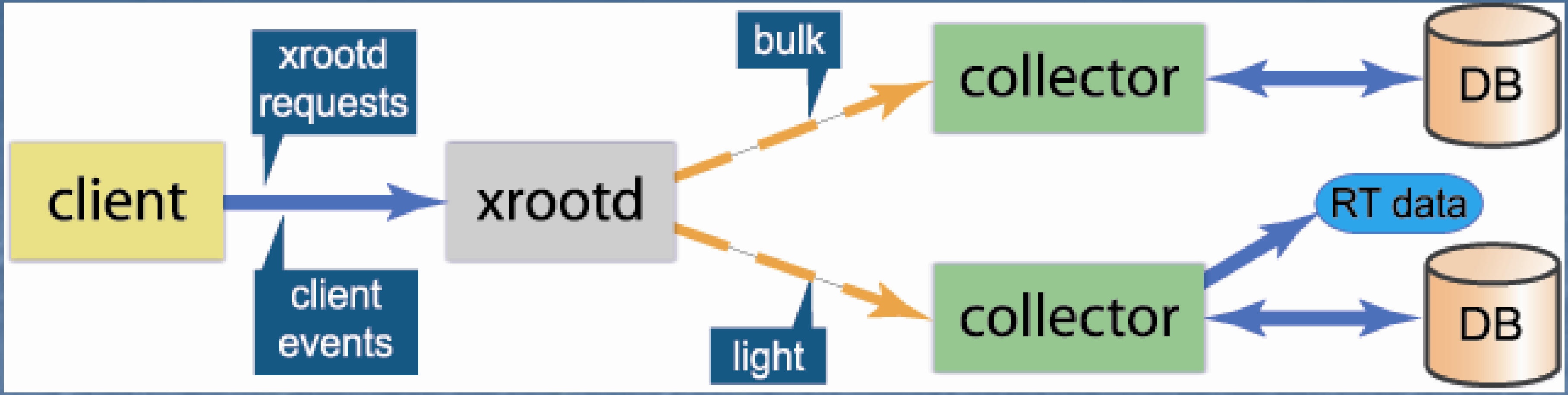
Use UDP datagrams

- 👍 Data servers insulated from monitoring. But
- 👎 Packets can get lost

Outsource client serialization

Low bounded resource usage

Use of time buckets



RT data

- **Start Session**

  sessionId, user, PId, client, server, start T

- **Staging**

  stageId, user, PId, client, file path, stage T, duration, server

- **Open File**

  fileId, user, PId, client, server, file path, open T

- **Close File**

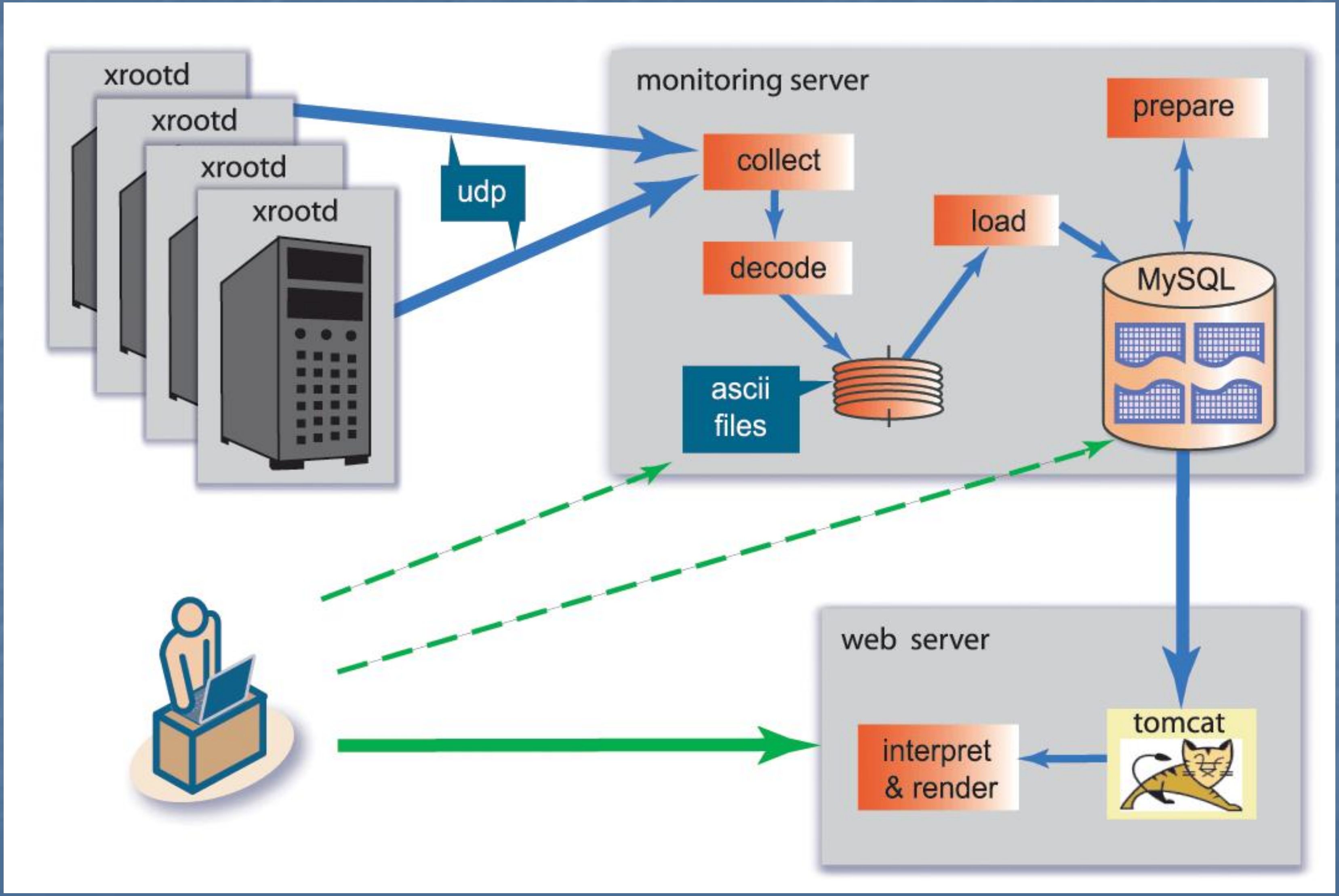
  fileId, bytes read, bytes written, close T

- **End Session**

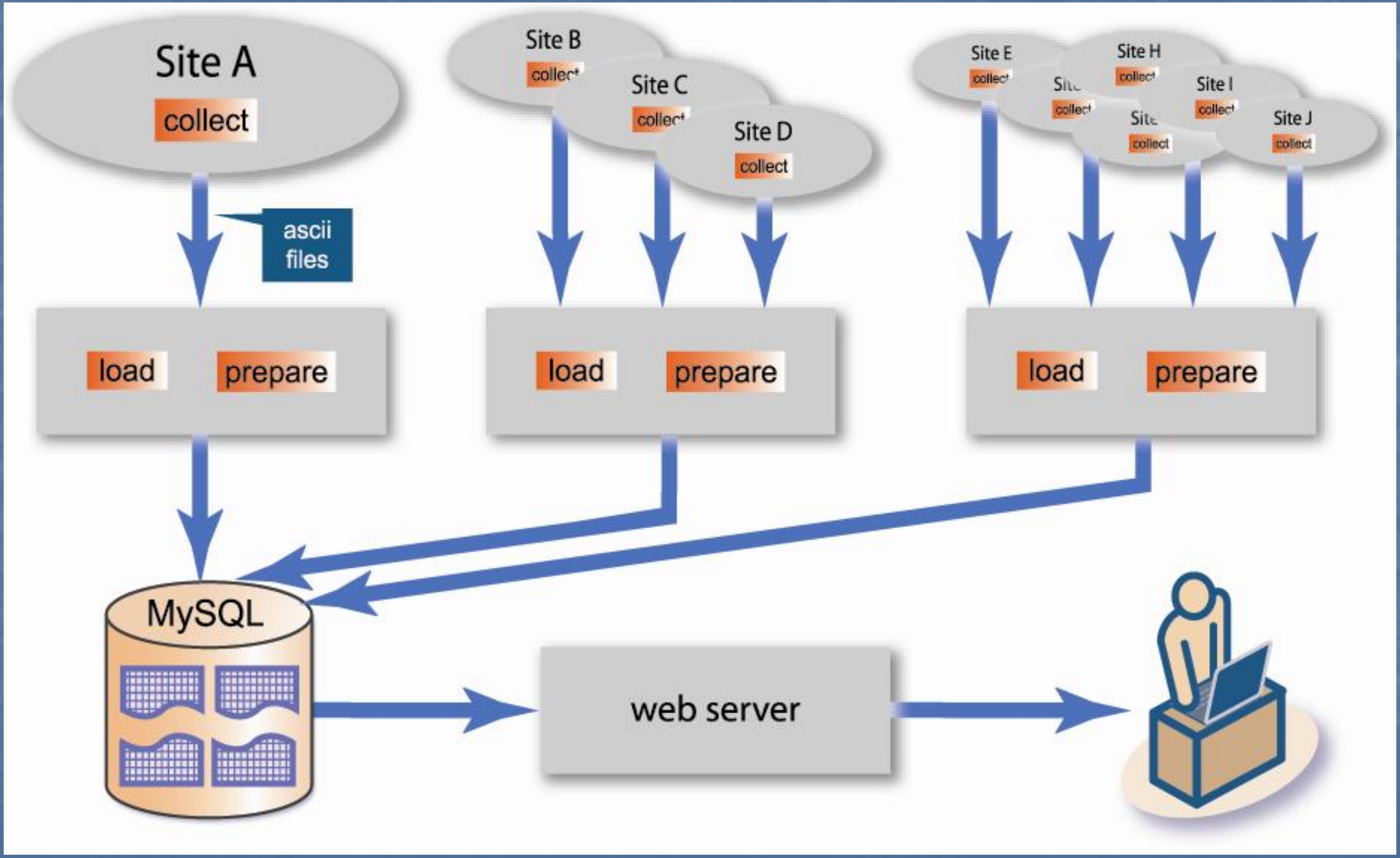
  sessionId, duration, end T

+ Xrootd restart time for each server

Single Site Monitoring



Multi-site Monitoring



COMPONENTS

- Collector/Decoder (C++)
- MySQL database (5.0)
- Database Applications (Perl, Perl DBI)
  - Create
  - Load
  - Prepare
  - Upgrade
  - Reload
  - Backup
- Web application (JSP3)
  - DB access via JDBC

- Data servers
  - xrootd enabled
- Database Server
  - Hosting DB & running DB application
- Web Server
  - Tomcat

For security reasons DB & Web servers on different hosts

Configuration File

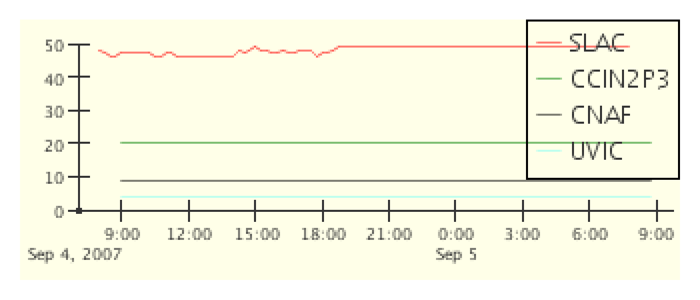
- dbName: xrdmon_kan_v005
- MySQLUser: xrdmon
- webUser: reader
- MySQLSocket: /tmp/mysql.sock
- baseDir: /u1/xrdmon/allSites
- ctrPort: 9931
- thisSite: SLAC
- fileType: dataType 100
- fileType: skim 500

- site: 1 SLAC PST8PDT 2005-06-13 00:00:00
- site: 2 RAL WET 2005-08-08 10:14:00
- site: 2 CCIN2P3 CET 2006-10-16 00:00:00
- site: 3 CNAF CET 2006-12-18 00:00:00
- site: 3 GRIDKA CET 2006-10-16 00:00:00
- site: 3 UVIC PST8PDT 2007-05-04 21:00:00
- backupInt: SLAC 1 DAY
- backupIntDef: 1 DAY

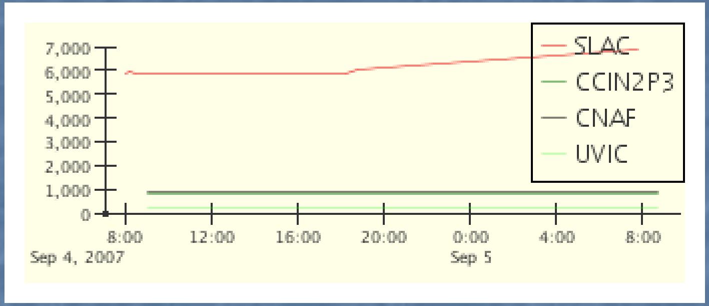
- fileCloseWaitTime: 10 MINUTE
- maxJobIdleTime: 15 MINUTE
- maxSessionIdleTime: 12 HOUR
- maxConnectTime: 70 DAY
- closeFileInt: 15 MINUTE
- closeIdleSessionInt: 1 HOUR
- closeLongSessionInt: 1 DAY
- nTopPerfRows: 20
- yearlyStats: ON
- allYearsStats: OFF

Basic View

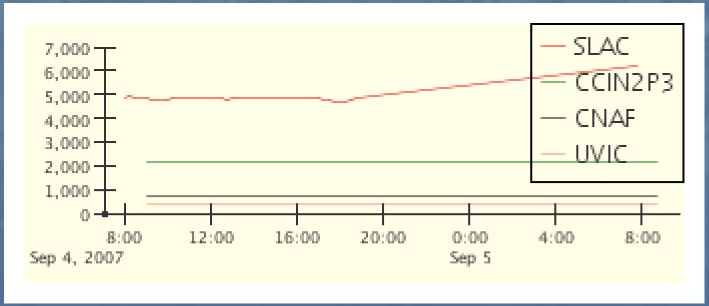
users



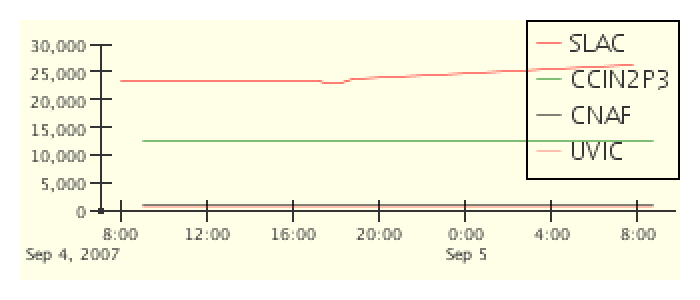
unique files



jobs



all files



Top Performers Table

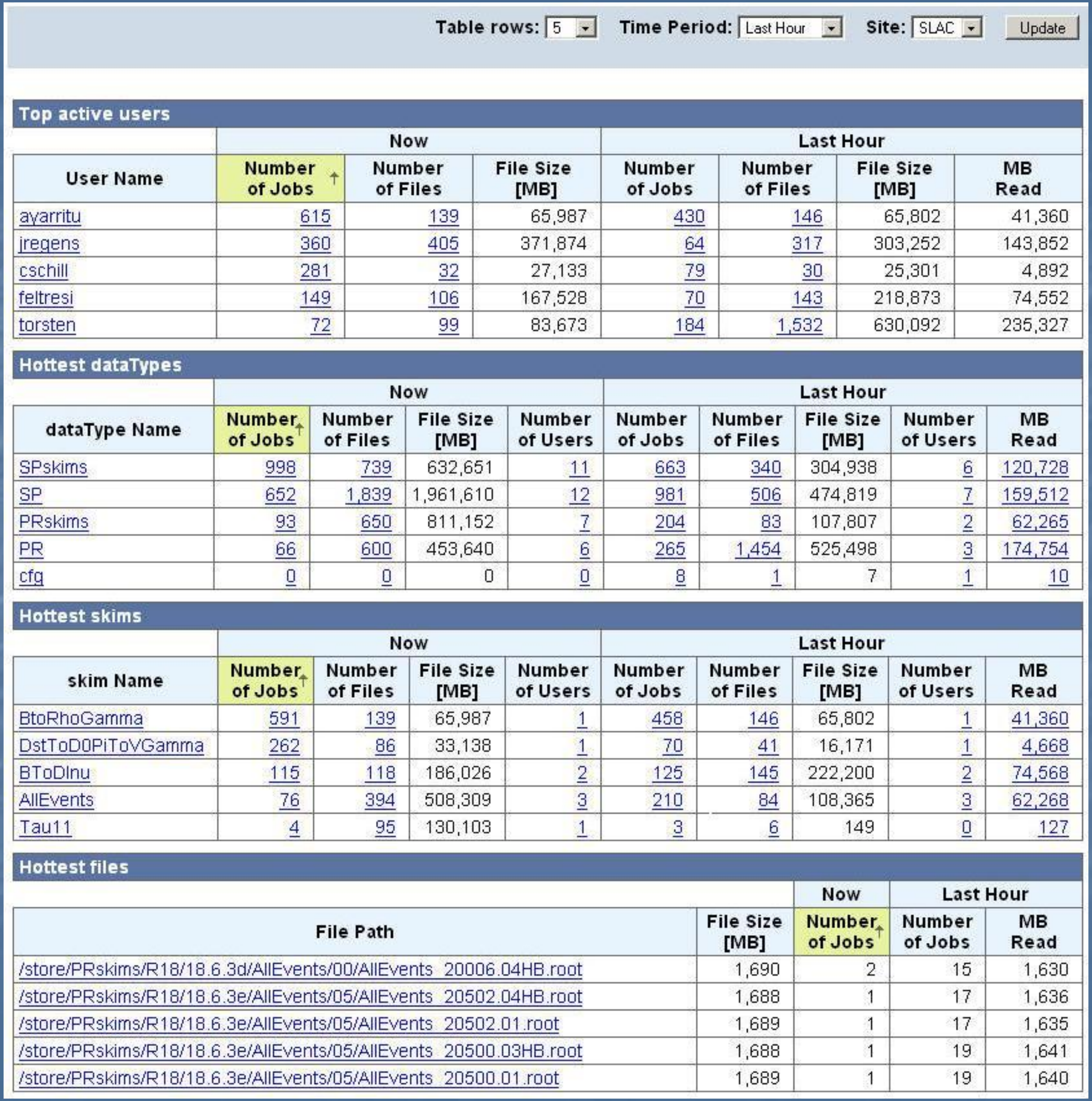


Table rows: 5 ▾ Time Period: Last Hour ▾ Site: SLAC ▾ Update

Top active users

	Now			Last Hour			
User Name	Number of Jobs ↑	Number of Files	File Size [MB]	Number of Jobs	Number of Files	File Size [MB]	MB Read
ayarritu	615	139	65,987	430	146	65,802	41,360
jregens	360	405	371,874	64	317	303,252	143,852
cschill	281	32	27,133	79	30	25,301	4,892
feltresi	149	106	167,528	70	143	218,873	74,552
torsten	72	99	83,673	184	1,532	630,092	235,327

Hottest dataTypes

	Now				Last Hour				
dataType Name	Number of Jobs ↑	Number of Files	File Size [MB]	Number of Users	Number of Jobs	Number of Files	File Size [MB]	Number of Users	MB Read
SPskims	998	739	632,651	11	663	340	304,938	6	120,728
SP	652	1,839	1,961,610	12	981	506	474,819	7	159,512
PRskims	93	650	811,152	7	204	83	107,807	2	62,265
PR	66	600	453,640	6	265	1,454	525,498	3	174,754
cfg	0	0	0	0	8	1	7	1	10

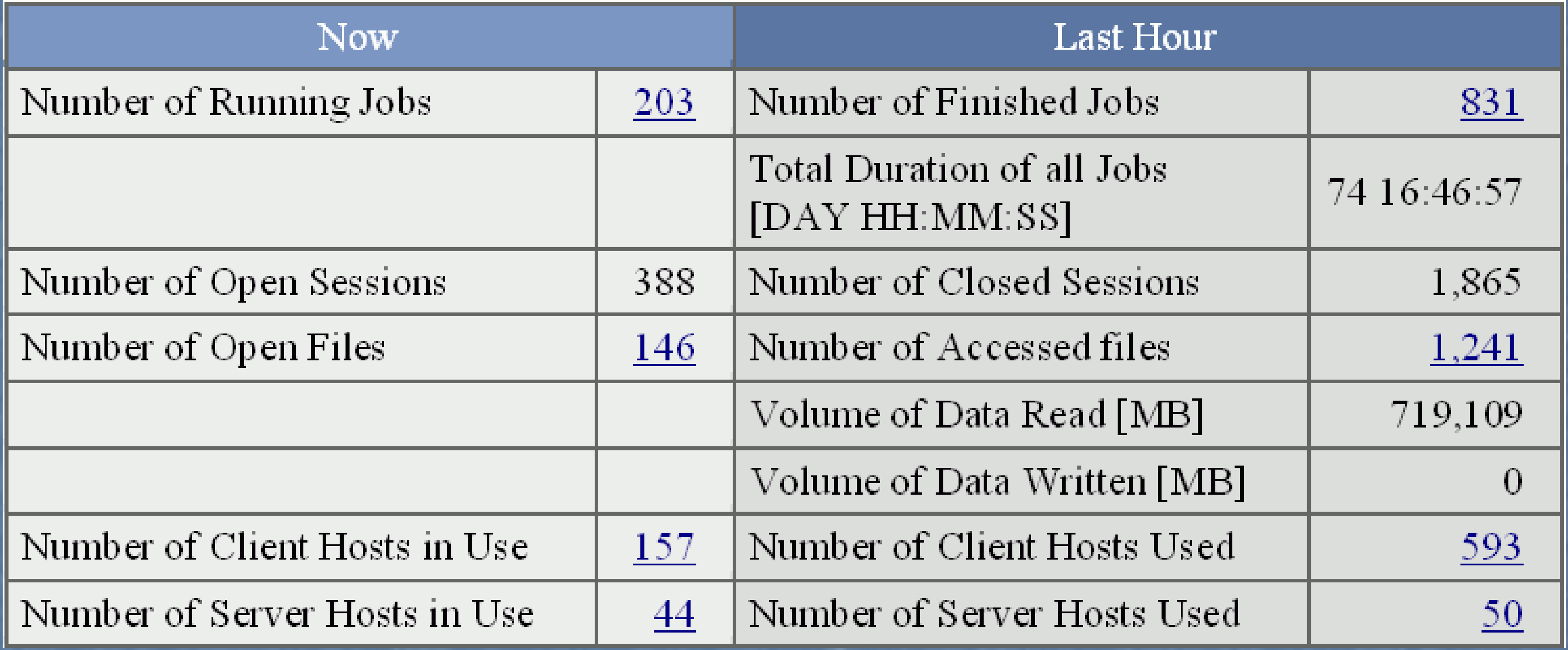
Hottest skims

	Now				Last Hour				
skim Name	Number of Jobs ↑	Number of Files	File Size [MB]	Number of Users	Number of Jobs	Number of Files	File Size [MB]	Number of Users	MB Read
BtoRhoGamma	591	139	65,987	1	458	146	65,802	1	41,360
DstToD0PiToVGamma	262	86	33,138	1	70	41	16,171	1	4,668
BToDlnu	115	118	186,026	2	125	145	222,200	2	74,568
AllEvents	76	394	508,309	3	210	84	108,365	3	62,268
Tau11	4	95	130,103	1	3	6	149	0	127

Hottest files

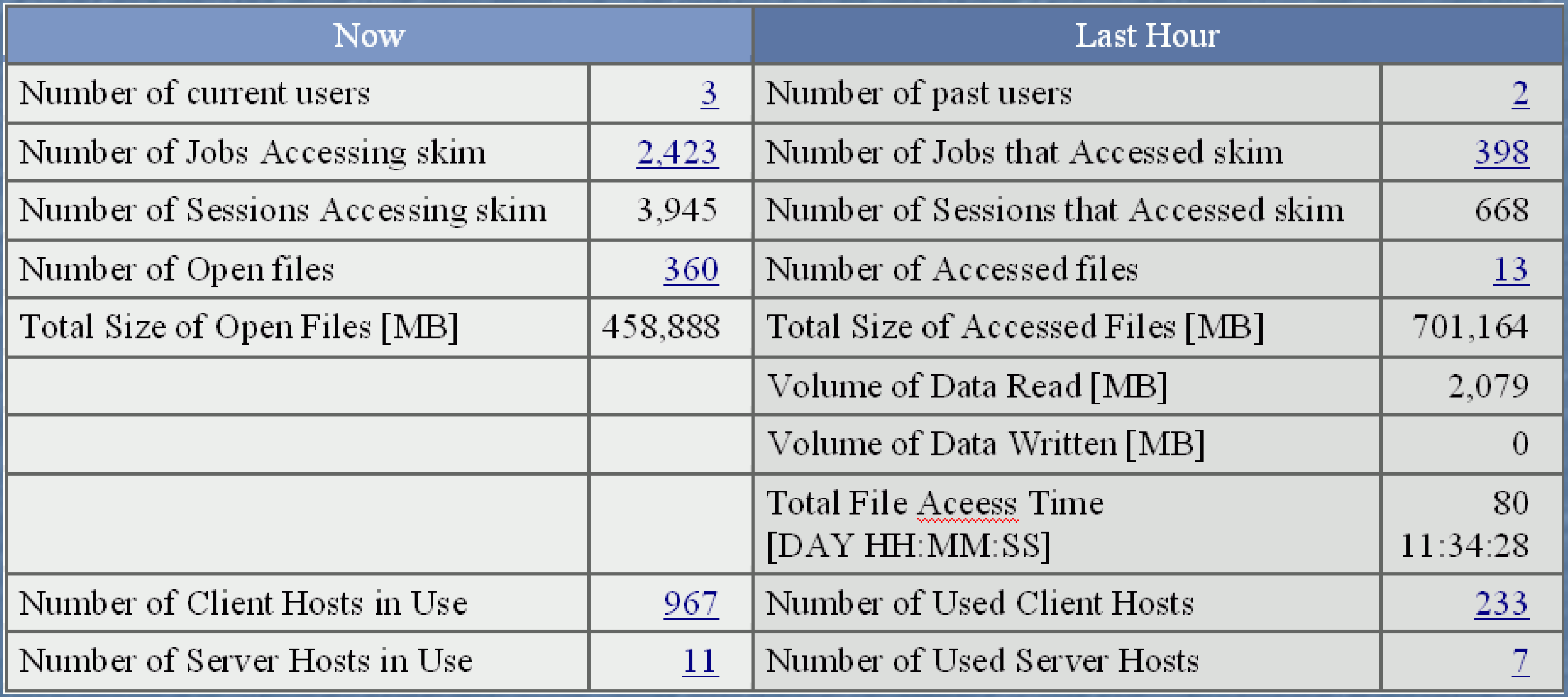
		Now	Last Hour	
File Path	File Size [MB]	Number of Jobs ↑	Number of Jobs	MB Read
/store/PRskims/R18/18.6.3d/AllEvents/00/AllEvents_20006.04HB.root	1,690	2	15	1,630
/store/PRskims/R18/18.6.3e/AllEvents/05/AllEvents_20502.04HB.root	1,688	1	17	1,636
/store/PRskims/R18/18.6.3e/AllEvents/05/AllEvents_20502.01.root	1,689	1	17	1,635
/store/PRskims/R18/18.6.3e/AllEvents/05/AllEvents_20500.03HB.root	1,688	1	19	1,641
/store/PRskims/R18/18.6.3e/AllEvents/05/AllEvents_20500.01.root	1,689	1	19	1,640

User Information



Now		Last Hour	
Number of Running Jobs	203	Number of Finished Jobs	831
		Total Duration of all Jobs [DAY HH:MM:SS]	74 16:46:57
Number of Open Sessions	388	Number of Closed Sessions	1,865
Number of Open Files	146	Number of Accessed files	1,241
		Volume of Data Read [MB]	719,109
		Volume of Data Written [MB]	0
Number of Client Hosts in Use	157	Number of Client Hosts Used	593
Number of Server Hosts in Use	44	Number of Server Hosts Used	50

Skim Information



Now		Last Hour	
Number of current users	3	Number of past users	2
Number of Jobs Accessing skim	2,423	Number of Jobs that Accessed skim	398
Number of Sessions Accessing skim	3,945	Number of Sessions that Accessed skim	668
Number of Open files	360	Number of Accessed files	13
Total Size of Open Files [MB]	458,888	Total Size of Accessed Files [MB]	701,164
		Volume of Data Read [MB]	2,079
		Volume of Data Written [MB]	0
		Total File Aceess Time [DAY HH:MM:SS]	80 11:34:28
Number of Client Hosts in Use	967	Number of Used Client Hosts	233
Number of Server Hosts in Use	11	Number of Used Server Hosts	7

Statistics (BaBar, September 2007)

- DB size: > 40 GB
- # tables: > 200
  - many with 10's of millions rows
  - Largest table > 132,000,000 rows
- # jobs recorded > 30,000,000
- At peak times
  - over 100 concurrent users
  - running 10's of thousands of jobs

Future Developments & Expansions

- DB and RT Data Backup
- File and User Filtering
- Staging Monitoring
  - Never enough disk to hold entire data sample
  - Disk uses power even when files are not accessed
- Fraction of Data Accessed
  - For each file type
  - In specific time intervals
  - …
- Multi Experiment Monitoring
  - Many experiments sharing computing resources