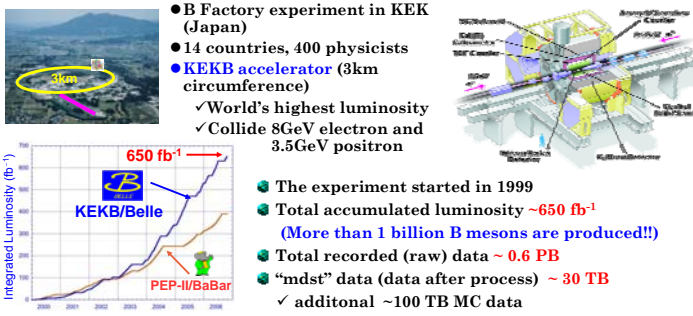


High Performance Data Analysis for Particle Physics using the Gfarm file system

Shohei Nishida, Nobuhiko Katayama, Ichiro Adachi (KEK)
Osamu Tatebe, Mitsuhsa Sato, Taisuke Boku, Akira Ukawa (Univ. of Tsukuba)

1. Belle Experiment



- B Factory experiment in KEK (Japan)
- 14 countries, 400 physicists
- KEKB accelerator (3km circumference)
 - ✓ World's highest luminosity
 - ✓ Collide 8GeV electron and 3.5GeV positron
- The experiment started in 1999
- Total accumulated luminosity ~650 fb⁻¹ (More than 1 billion B mesons are produced!!)
- Total recorded (raw) data ~ 0.6 PB
- "mdst" data (data after process) ~ 30 TB
 - ✓ additional ~100 TB MC data
- Users read "mdst" data for analysis

In the present computing system in Belle:

- ▶ Data are stored under ~40 file servers (FS) :storage with 1PB disk + 3.5PB tape
- ▶ 1100 computing servers (CS) for analysis, simulations...
- ▶ Data are transferred from FS to CS using Belle home grown TCP/socket application
- ▶ It takes 1~3 weeks for one physicist to read all the 30TB data

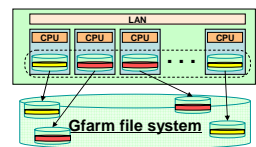
Computing Servers in B Factory Computer System

- 1140 compute nodes
 - ▶ DELL PowerEdge 1855 blade server
 - ▶ 3.6GHz Dual Xeon
 - ▶ 1GB Memory, 72GB RAID-1
- 80 login nodes
 - ▶ Gigabit Ethernet
 - ▶ 24 EdgeIron 48GS
 - ▶ 2 BigIron RX-16
 - ▶ Bisection 9.8 GB/s
- Total: 45662 SPECint2000 Rate

1 enclosure = 10 nodes / 7U space
1 rack = 50 nodes

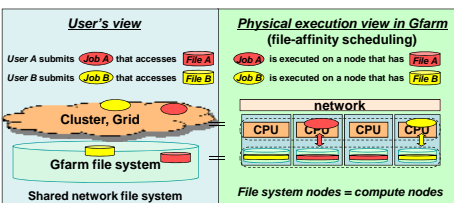
2. Gfarm

- Commodity-based distributed file system that federates local disks of compute nodes
- It can be shared among all cluster nodes and clients
 - ✓ Just mount it as if it were high-performance NFS
- It provides scalable I/O performance w.r.t. the number of parallel processes and users
- It supports fault tolerance and avoids access concentration by automatic replica selection



- Files can be shared among all nodes and clients
- Physically, it may be replicated and stored on any file system node
- Applications can access it regardless of its location
- File system nodes can be distributed

Scalable I/O Performance



- Do not separate storage and CPU (SAN not necessary)
- Move and execute program instead of moving large-scale data
- exploiting local I/O is a key for scalable I/O performance

- libgfarm - Gfarm client library
 - ▶ Gfarm API
- Metadata server, metadata cache servers
 - ▶ Namespace, replica catalog, host information, process information
- gfsd - I/O server
 - ▶ file access

<http://datafarm.apgrid.org/>

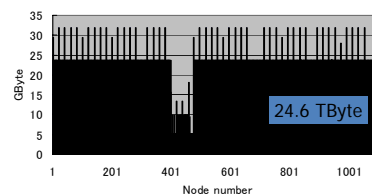
3. Challenge

Use Gfarm file system for Belle analysis

- Scalability of Gfarm File System up to 1000 nodes
 - ▶ Scalable capacity federating local disk of 1112 nodes
 - 24 GByte x 1112 nodes = 26 TByte
 - ▶ Scalable disk I/O bandwidth up to 1112 nodes
 - 48 MB/sec x 1112 nodes = 52 GB/sec
 - Speed up of KEKB/Belle data analysis
 - ▶ Read 25 TB of "mdst" (reconstructed) real data taken by the KEKB B factory within 10 minutes
 - For now, it takes 1 - 3 weeks
 - ▶ Search for the Direct CP asymmetry in $b \rightarrow s \gamma$ decays
 - It may provide clues about physics beyond the standard model
- Goal: 1/1000 Analysis time with Gfarm

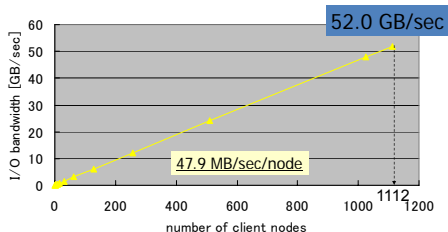
- 1112 file system nodes
 - ▶ = 1112 compute nodes
 - ▶ 23.9 GB x 1061 + 29.5 GB x 8 + 32.0 GB x 43 = 26.3 TB
- 1 metadata server
- 3 metadata cache server
 - ▶ one for 400 clients
- All hardware is commodity
- All software is open source

24.6 TB of reconstructed data is stored on local disks of 1112 compute nodes



4. Measurement

Read I/O Bench mark

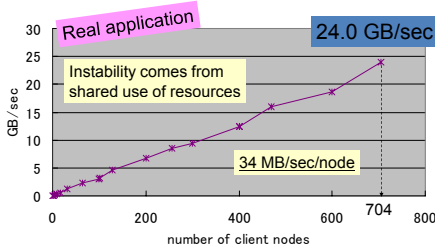


We succeeded in reading all the reconstructed data within 10 minutes

Completely scalable bandwidth improvement, and 100% of peak disk I/O bandwidth are obtained

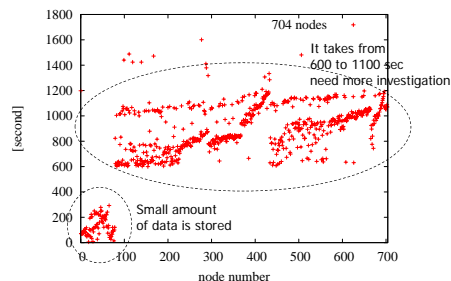
Running Belle Analysis Software

- Read "mdst" data and skim (select) useful events for $b \rightarrow s \gamma$ analysis.
- Output file is index file (event number lists)



Scalable data rate improvement is observed

Break down of skimming time



5. Conclusion

- Scalability of Commodity-based Gfarm File System up to 1000 nodes is shown
 - ✓ Capacity 26TB, Bandwidth 52.0 GB/sec
 - ✓ Novel file system approach enables such scalability.
- Read 24.6 TB of "mdst" data within 10 minutes (52.0 GB/sec)
 - ✓ 24.0 GB/sec
 - ✓ 3,000 times speedup for disk I/O

Our team is the winner at the Storage Challenge in the SC06 conference:

