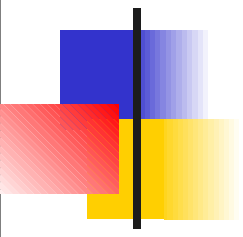
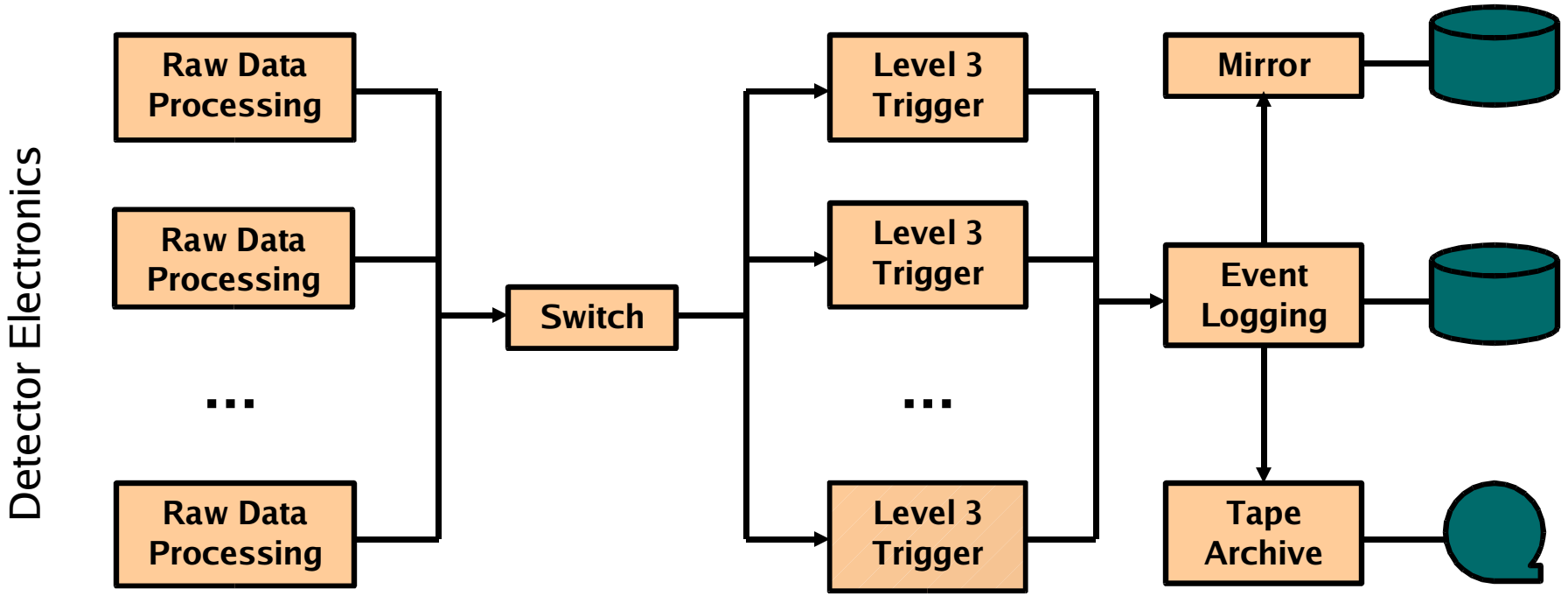


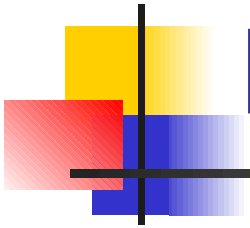
# A New Scalable Multi-Node Event Logging System for BaBar

James A. Hamilton
Steffen Luitz

For the BaBar Computing Group

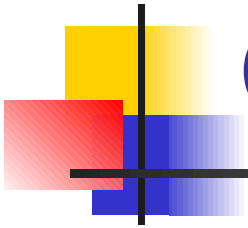
# Original Structure





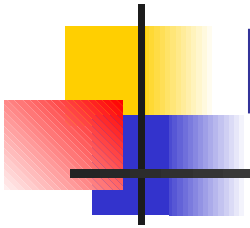
# Performance Requirements

- Current
  - *30 kilobyte events X 300 events/sec. = 9 megabytes/sec.*
  - *Server limited to ~ 10 megabytes/sec.*
- Projected by 2007
  - *75 kilobyte events X 500 events/sec. = 37.5 megabytes/sec.*



# Options

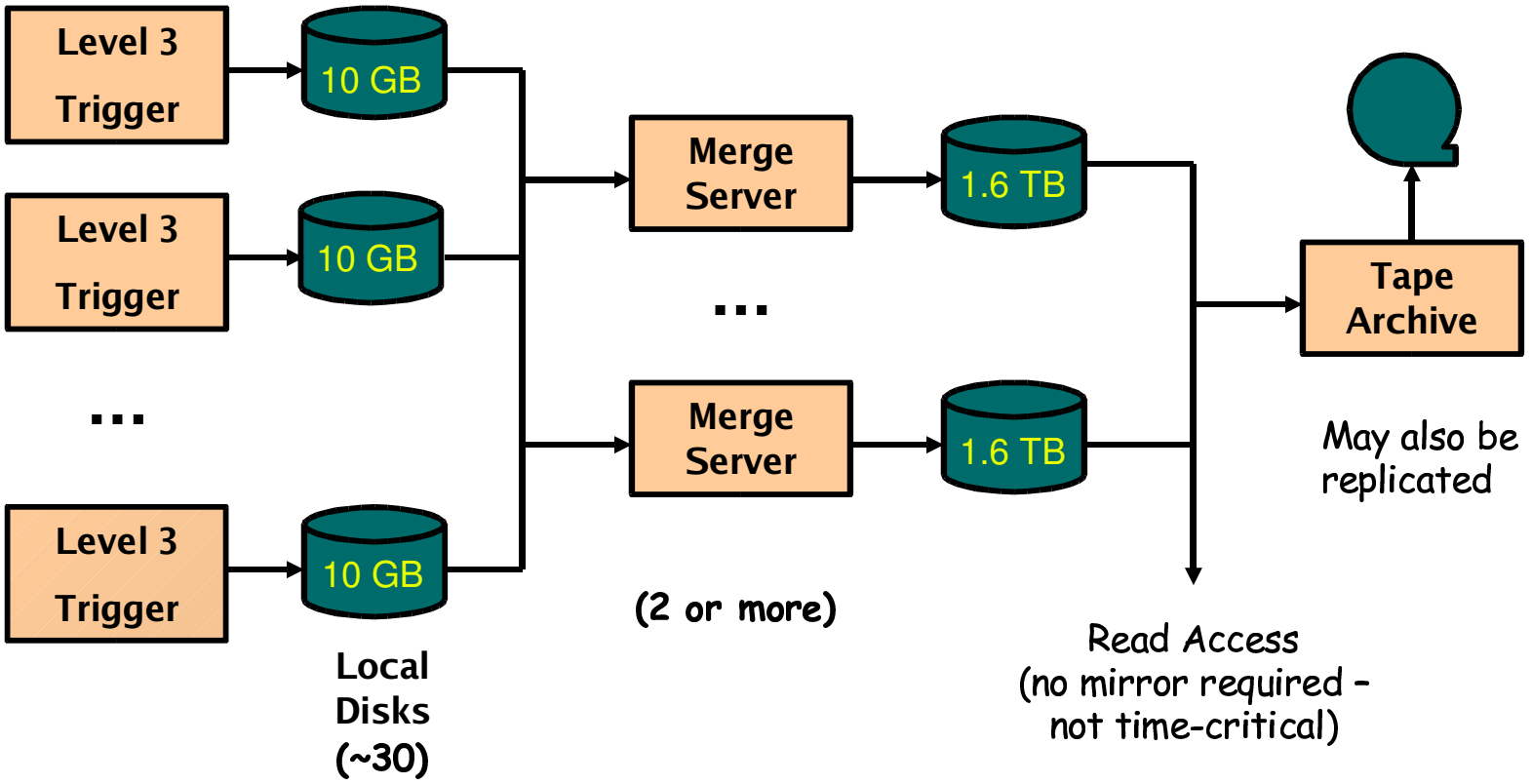
- Could do a simple HW/SW upgrade
- But a new architecture can also get:
  - *Scalability in case we guess wrong*
  - *Fault tolerance (RAID array not enough because of long disk reconstruction time)*
  - *High burst rate to allow passthrough runs*
- So we chose the new architecture

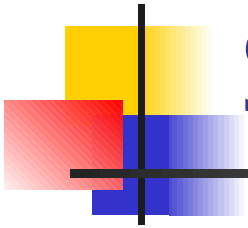


# New Architecture

- Take advantage of multiple level 3 systems (currently 30)
- Write data in parallel to local disks on level 3 systems
- Merge to file server at end of run
  - *BaBar B-factory runs continuously with ~1 hour runs*
  - *Uses trickle injection, so data-taking never stops*
- Multiple file servers for finished runs

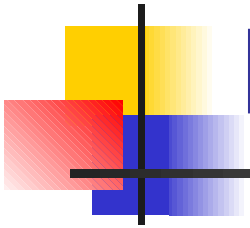
# Data Paths


Level 3 Trigger
Level 3 Trigger
...
Level 3 Trigger
10 GB
10 GB
10 GB
Local Disks (~30)
Merge Server
...
Merge Server
(2 or more)
1.6 TB
1.6 TB
Tape Archive
May also be replicated
Read Access
(no mirror required - not time-critical)



# Server Performance

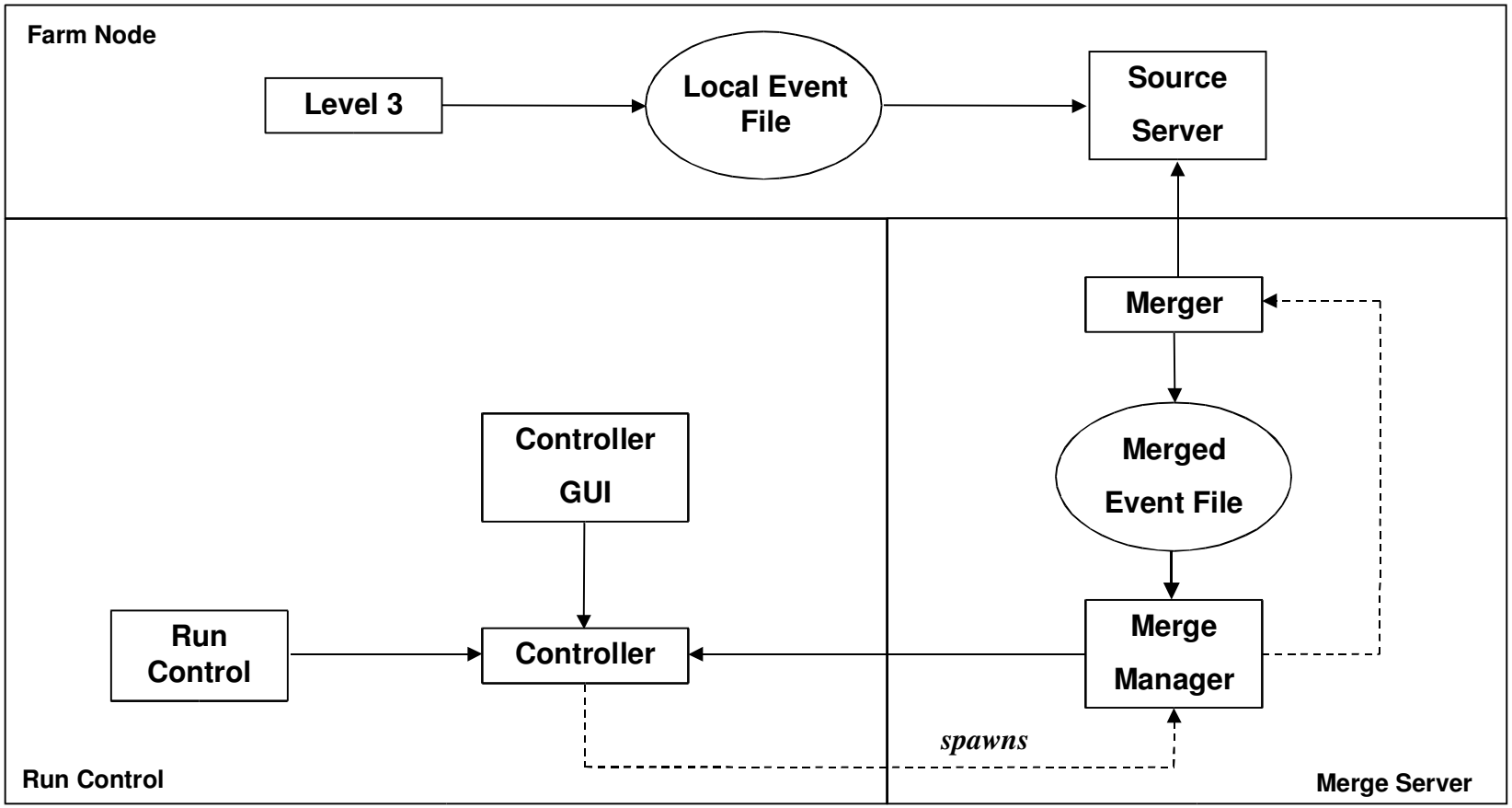
- Each merge server is equipped with:
  - *1.6 terabyte raid disk array at ~100 megabytes/sec.*
  - *Gigabit ethernet*
  - *Dual 3GHz P4 running Linux & XFS*
- Merge operations average 80 megabytes/sec. transfer rate
  - *Over 1000 events/sec. At 75 kilobytes per event*
- Use second server for fault tolerance



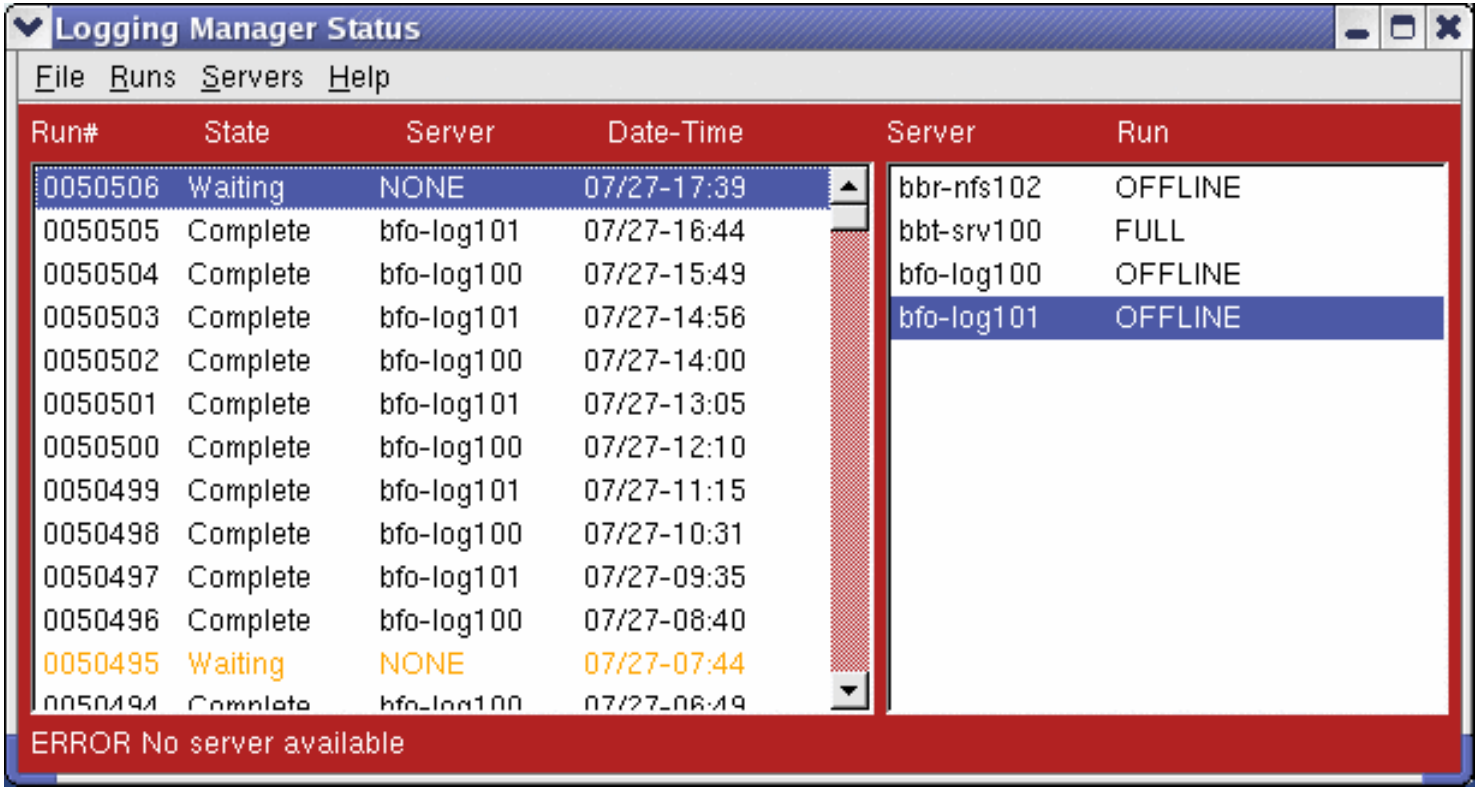
# Level 3 System Performance

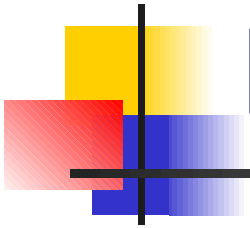
- Each level 3 system has:
  - *Two 18 gigabyte mirrored disks at ~50 megabytes/sec.*
    - Leaves ~10 gigabytes for event data – enough for ~10 hours at current rates
    - Need larger disks here already
  - *Gigabit ethernet to dataflow*
  - *100 megabit ethernet to merge servers*
- Each merger takes 2.6 megabytes/sec. (80/30)
  - *5 % of disk bandwidth*
  - *21 % of network bandwidth*
- ~20,000 events/sec. burst, at 75 kilobytes/event
  - *Have done 3,000 events/sec. in production*

# Component Architecture



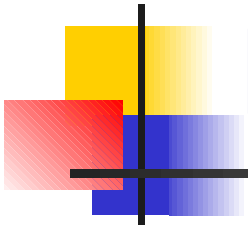
# Controller GUI





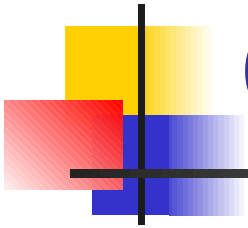
# Failure Modes

- Loss of network connection
- Unrecoverable single disk error
- Processor failure
- Component crash or restart
- Must recover from these without loss of data



# Merge Failure Recovery

- Repair
  - *Replace failing drive (recover via RAID) or processor*
  - *Continue taking data during repair*
- Retry
  - *Source data never deleted until merge succeeds*
  - *Manual and automatic*
- Scavenge
  - *Controller queries source servers and merge manager*
- Redirect
  - *Sources can be renamed or omitted*



# Conclusions

- We have developed a new event logging system that is:
  - *Scalable – meets projected throughput requirements for lifetime of BaBar*
  - *Fault tolerant*
  - *Supports high burst-rate data taking*
- In successful production use for last two weeks of Run 4. Will continue in Run 5.

# Scalability Timelines