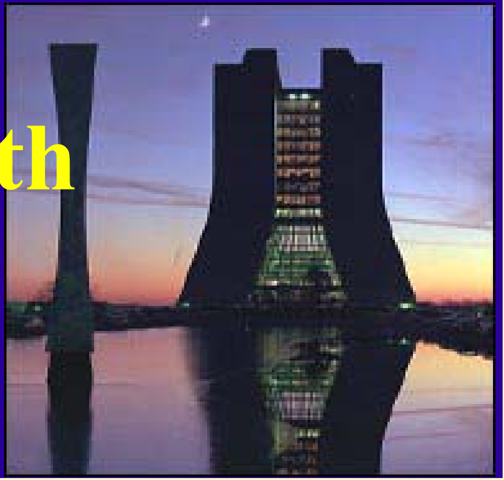
”The CMS Tier 1 Computing Center at Fermilab"

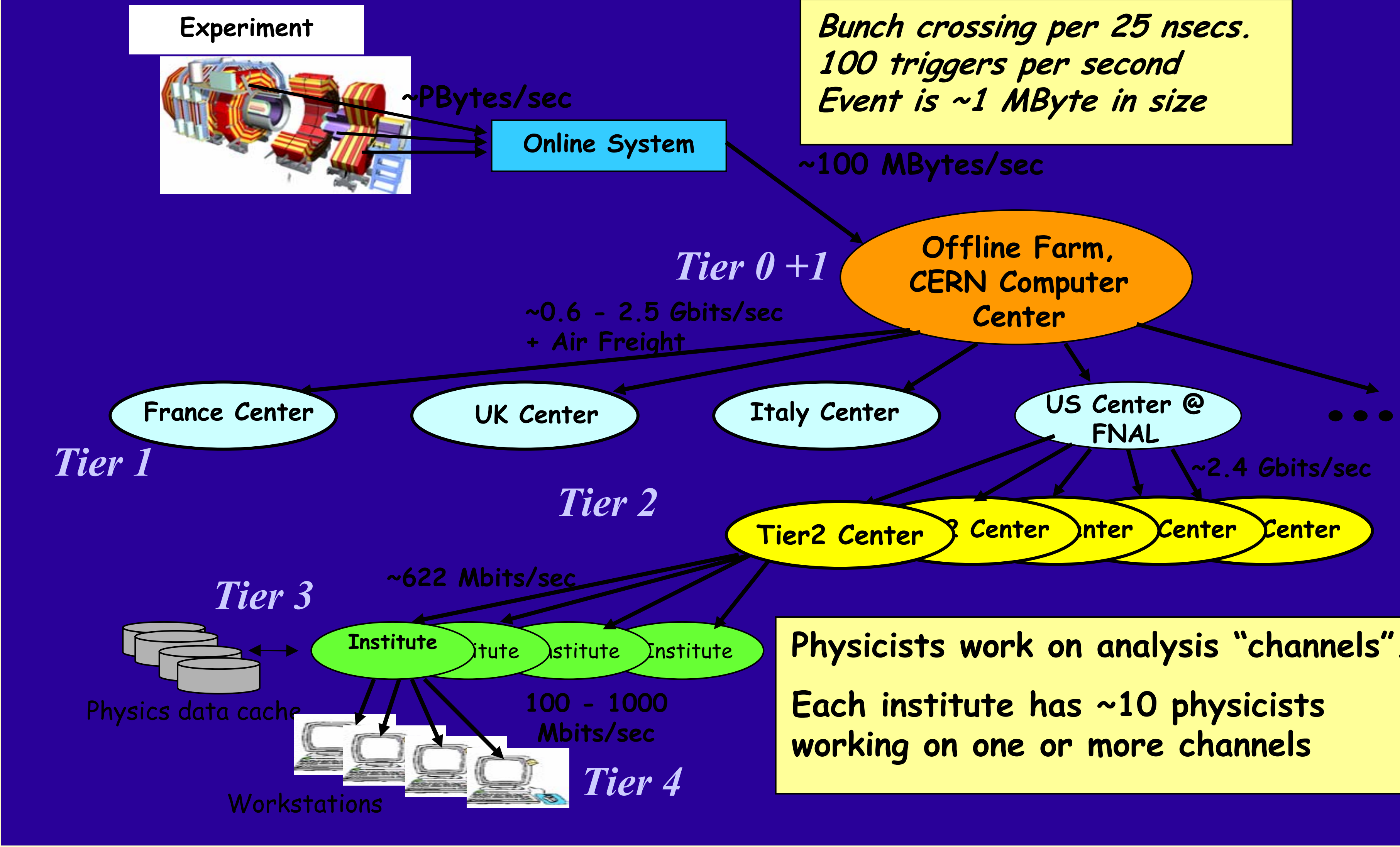
Hans Wenzel Fermilab

▪

- ❑ The big picture (how do we fit in).
- ❑What do we do.
- ❑ What cms computing resources are Currently deployed at FNAL.
- ❑ First benchmarking results with dCache
- ❑ Plans for the near term future



US and Worldwide Data Grid for CMS



Experiment

~PBytes/sec

Online System

~100 MBytes/sec

Tier 0 +1

Offline Farm, CERN Computer Center

~0.6 - 2.5 Gbits/sec + Air Freight

France Center

UK Center

Italy Center

US Center @ FNAL

• • •

Tier 1

~2.4 Gbits/sec

Tier 2

Tier2 Center

Tier 3

~622 Mbits/sec

Institute

Physics data cache

100 - 1000 Mbits/sec

Workstations

Tier 4

Bunch crossing per 25 nsecs.
100 triggers per second
Event is ~1 MByte in size

Physicists work on analysis "channels".

Each institute has ~10 physicists working on one or more channels

MONARC defines several levels (Tiers) of Regional Center; This model supports a highly distributed
infrastructure both for technical reasons (e.g., to place computational
and data resources near to demand) and for strategic motives (e.g., to
leverage existing expertise and technology investments).

- Tier 1 center having roughly 20% of the capacity of CERN for a single experiment
- Each Tier 2 site will contain 20-25% of the computational capacity of the Fermilab Tier 1 center; so five centers would have approximately the same combined CPU capacity as the Tier 1 facility.
- The Tier2 sites will be sited in different regions of the US, and located at universities which have significant existing computing infrastructure and good connectivity to regional networks. Thus sites can minimize costs by leveraging existing facilities and support personnel, as well as bringing in additional resources. The responsibilities of the Tier 2 sides include:

(1) simulation (including reconstruction of simulated events),

(2) user analysis

(3) testing and other services in support of distributed analysis and the CMS data grid. The CPU anddisk capacities reflect the fact that almost all simulations are performed on Tier 2 equipment and each center has a share of 20% of US physicists for analysis.

The mission of the “S&C Project”

- To provide the software and computing resources needed to enable US physicists to fully participate in the physics program of CMS

- Allow US physicists to play key roles and exert an appropriate level of leadership in all stages of computing related activities …
  - From software infrastructure to reconstruction to extraction of physics results
  - From their home institutions, as well as at CERN or FNAL

Introduction

I. computing at the CMS Tier 1 center at FNAL provides:

II. Monte Carlo Production (Trigger + physics TDR) in distributed environment.

III. Host and serve the data, Mass storage

IV. Provide computing and development platform for physicist (resources, code, disk, help, tutorials,.....)

V. Evaluate new hardware, software solutions

VI. Active development

VII. The scope and complexity of a CMS tier one center is very comparable to the computing needs of the ongoing run II experiments CDF and D0. So at Fermilab we have the unique opportunity to look over their shoulders to see what works for them (and what doesn't). The CDF approach of using farms for user computing going away from SMP machines which don't provide a lot of bang for the buck. Also the approach of integrating the desktop is exactly the CMS approach.

Our Web sites, Info about the tools we are using (most of them have been already mentioned several times during this workshop)

I. Monitoring page, links to tools and scripts http://computing.fnal.gov/cms/Monitor/cms_production.html

II. The ganglia low level monitoring system http://gyoza8.fnal.gov/

III. **Department web site:** http://computing.fnal.gov/cms

IV. The Batch system we use is FBSNG which has been especially developed for farms. See Igor's talk http://gyoza8.fnal.gov/cgi-bin/fbsng/fbswww/fbswww

V. The online dCache page: http://gyoza7.fnal.gov:443

VI. The dCache page at DESY http://dcache.desy.de/summaryIndex.html

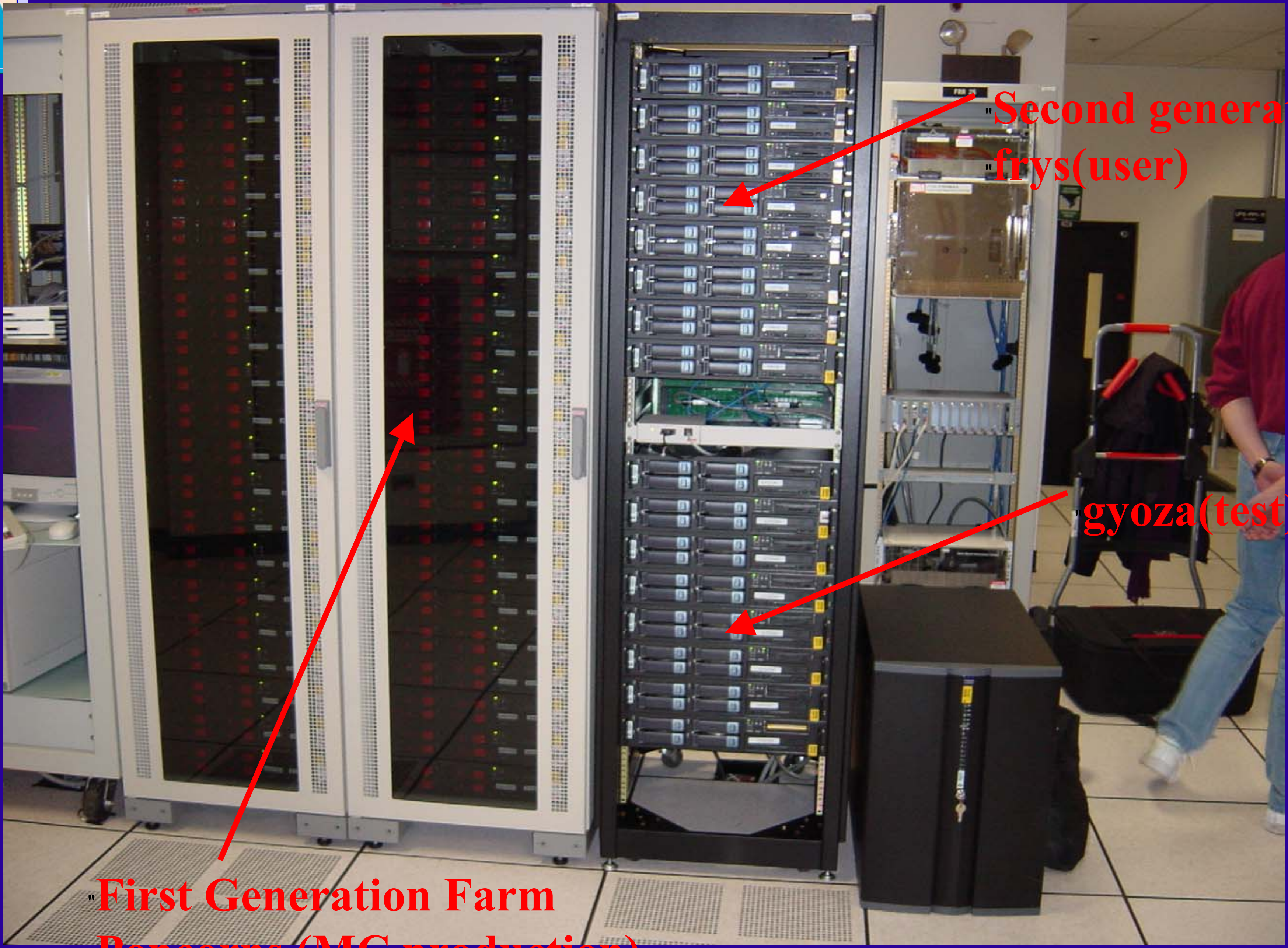
VII. http://computing.fnal.gov/cms/hans/massstorage/

Hardware selection +Purchasing

- Evaluate new hardware (CPU, motherboards…) criteria: CPU performance, IO performance, memory/networking bandwidth, cooling, power consumption, performance of physics application, compatibility with Fermi Linux etc…
- Request for bids (select vendor)
- 4 week burn in period during the time we exercise and monitor. We only accept after the farm passes the burn in period.
- Currently (only) 3 generation of farms on the floor.

Status of Upgrades

- 65 dual amd athlon 1900+ nodes are installed and will now start to undergo a 4 weeks burn in acceptance phase (probably >20 nodes for user computing)
- 7 dCache linux nodes: hardware has been upgraded (SCSI system disk). We are in the process of upgrading the software (Kernel 2.4.18, XFS File system). System is usable during the upgrades (see tests). 8-12 servers by the end of the year
- We will get a 3TB system in from zambeel for evaluation this week.
- Faster higher capacity tape drives stk 9940b
- Better connectivity between cms computing and e.g. mass storage.

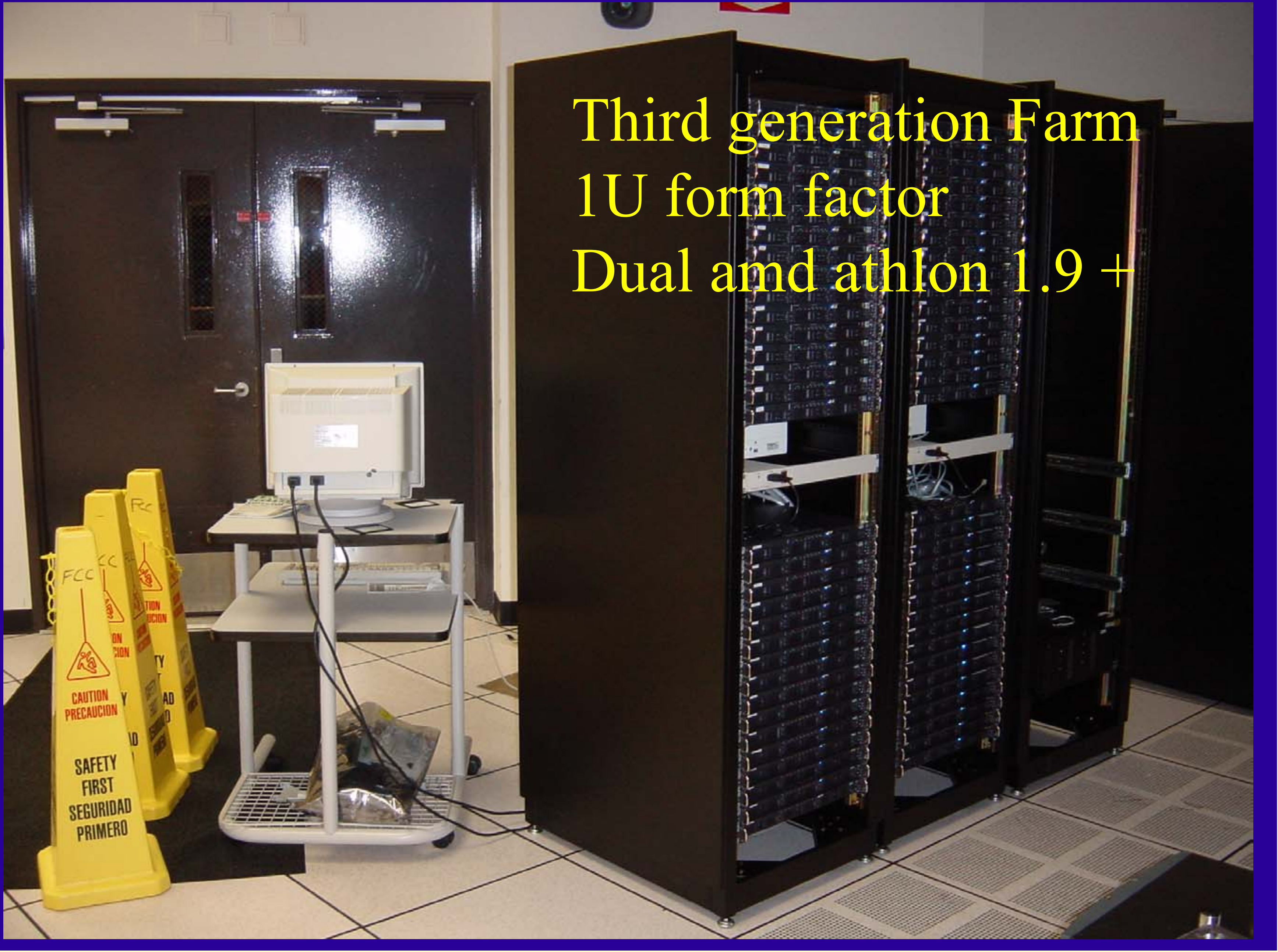


Second generation frys(user)

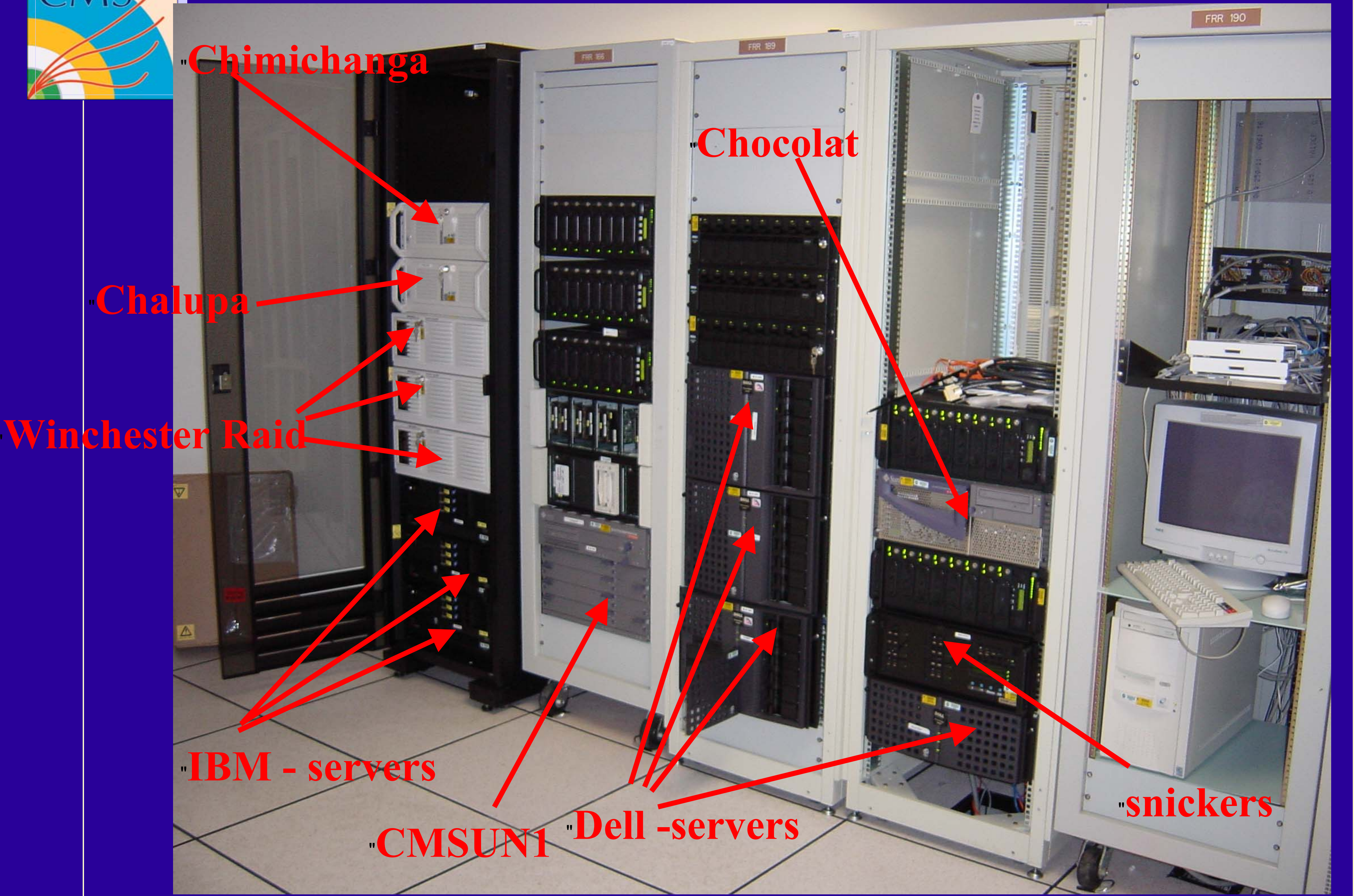
gyoza(test)

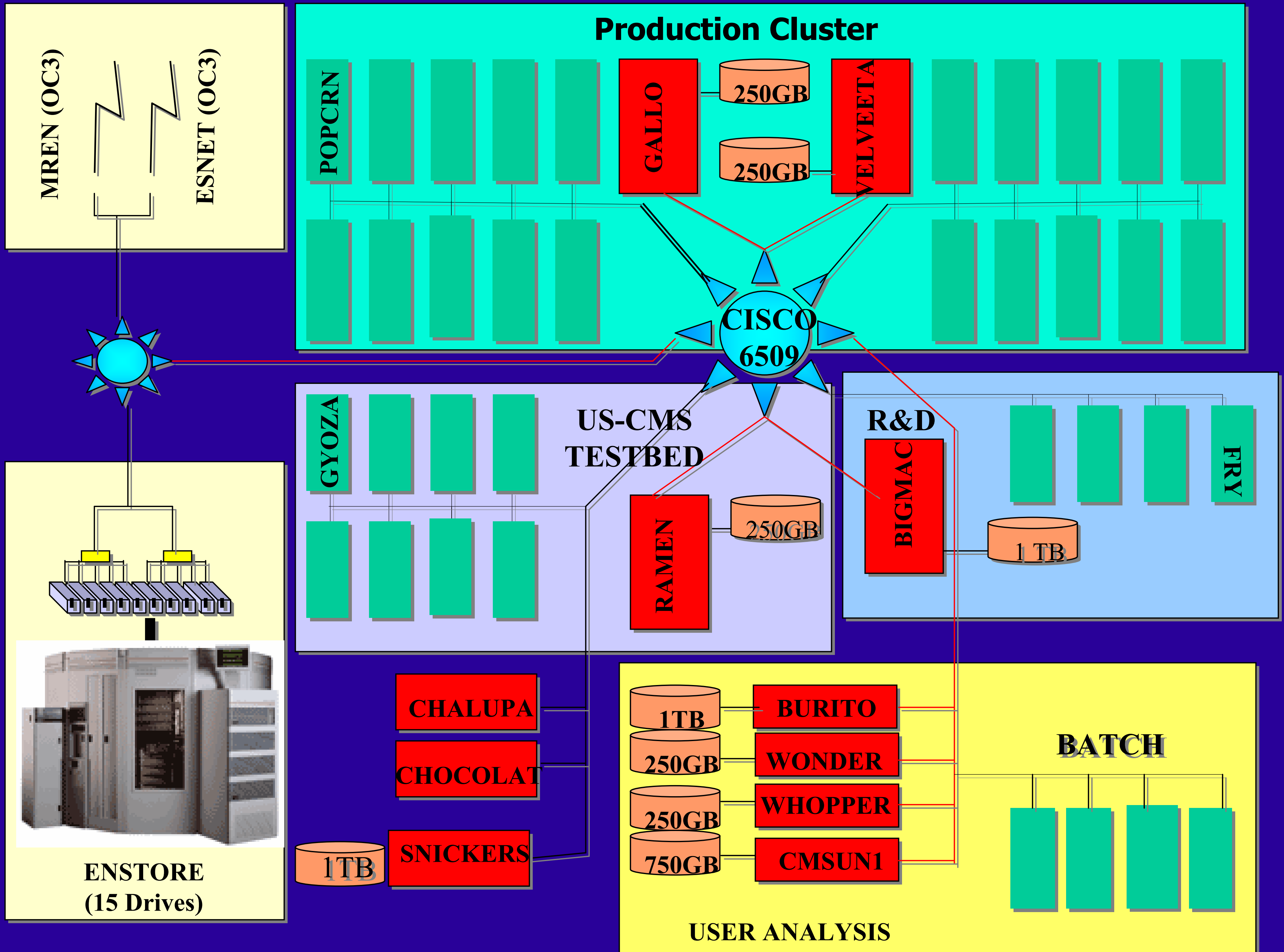
First Generation Farm Popcorns (MC production)

The new farm



Third generation Farm
1U form factor
Dual amd athlon 1.9 +


Chimichanga
Chalupa
Winchester Raid
Chocolat
IBM - servers
CMSUN1
Dell -servers
snickers
FRR 166
FRR 189
FRR 190



MREN (OC3)

ESNET (OC3)

Production Cluster

POPCRN

GALLO

250GB

250GB

VELVEETA

CISCO 6509

US-CMS TESTBED

GYOZA

RAMEN

250GB

R&D

BIGMAC

1 TB

FRY

ENSTORE (15 Drives)

CHALUPA

CHOCOLAT

SNICKERS

1TB

1TB

BURITO

250GB

WONDER

250GB

WHOPPER

750GB

CMSUN1

BATCH

USER ANALYSIS

What's available for the User at FNAL

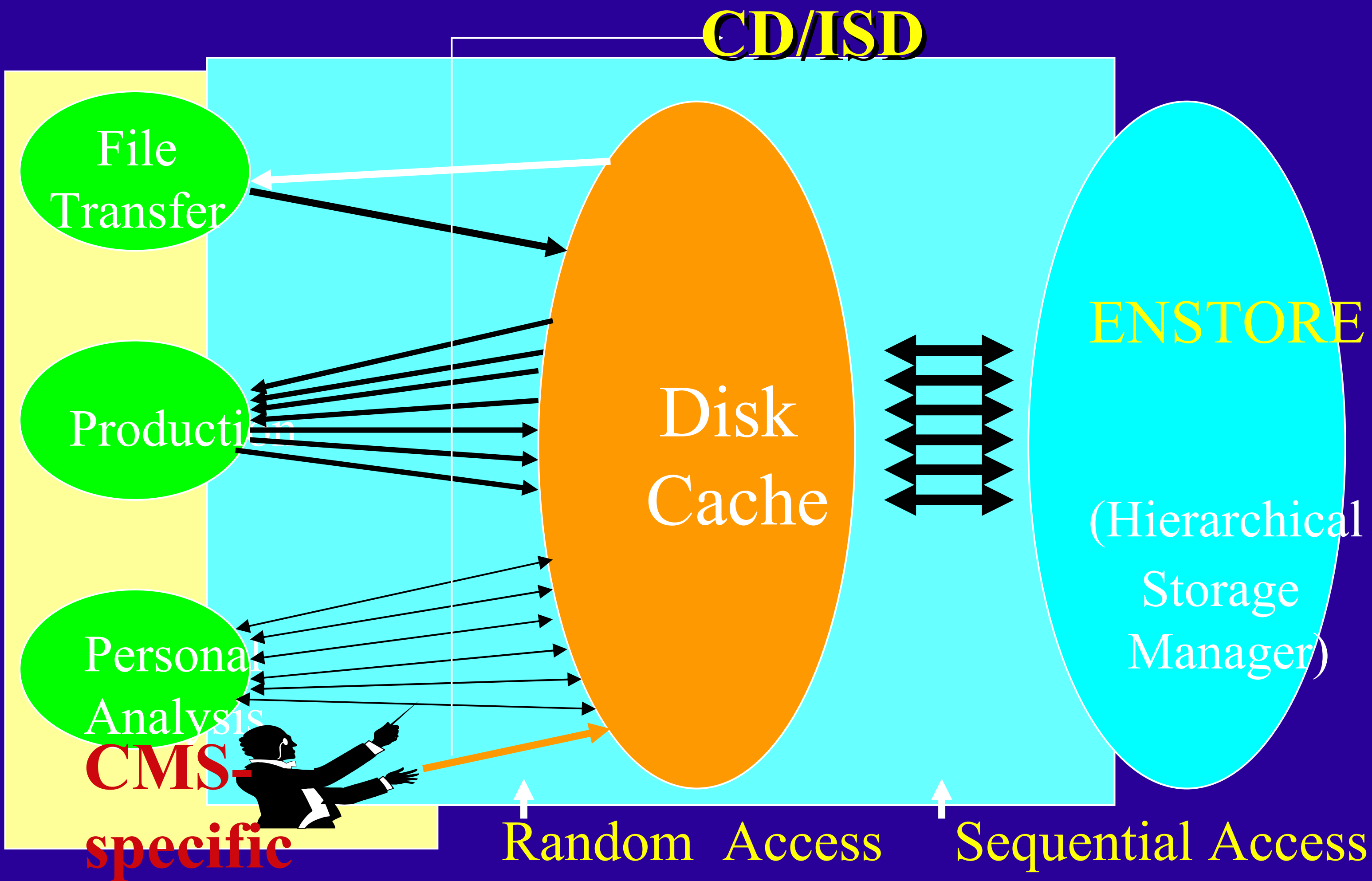
■

/ linux 4 way servers: wonder, burrito, whopper, nfs cross mounted /data disks (DAS), FBSNG batch system (8 CPU’s) attached to whopper. User needs to get kerberos principle matched to special batch principal. Home areas in AFS.

/ Cmsun1: 8 way sun smp machine

/ Plan to use the farm of linux node for interactive computing.

The current system consists of :
5 x 1.2 TB (Linux) read pools 1X 0.1 TB (write pool)
1sun server +1/4 TB raid array as write pool. We have additional 2 servers for R&D and funding for more (>5).



What do we expect from dCache?

- ❑ making a multi-terabyte server farm look like one coherent and homogeneous storage system.
- ❑ Rate adaptation between the application and the tertiary storage resources.
- ❑ Optimized usage of expensive tape robot systems and drives by coordinated read and write requests. Use dccp command instead of encp!
- ❑ No explicit staging is necessary to access the data (but prestaging possible and in some cases desirable).
- ❑ The data access method is unique independent of where the data resides.
- ❑ High performance and fault tolerant transport protocol between applications and data servers
- ❑ Fault tolerant, no specialized servers which can cause severe downtime when
- ❑ Can be accessed directly from your application (e.g. root TDCacheFile class).

Linux dCache node:

Basically same configuration as CDF.(but 120 WD disks). Important Kernel 2.4.18 to avoid memory management problems

XFS filesystem: we found it's the only one that scales still

Delivers performance when File system is full

Add SCSI system disk

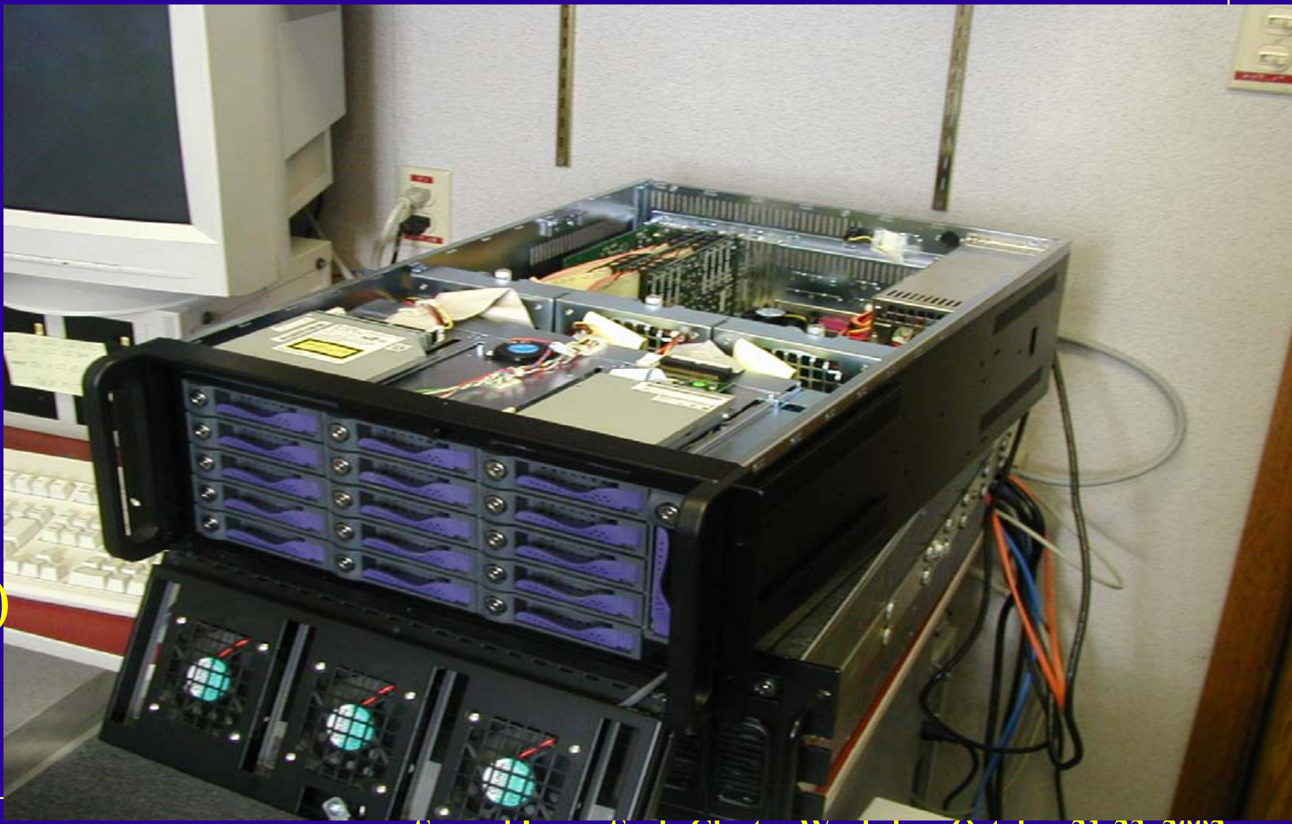
Need server specific Linux distribution!!! (same true for user machines)

Next generation:

Xeon based, PCIX bus

Large capacity disks

Dual System disk (raid1)



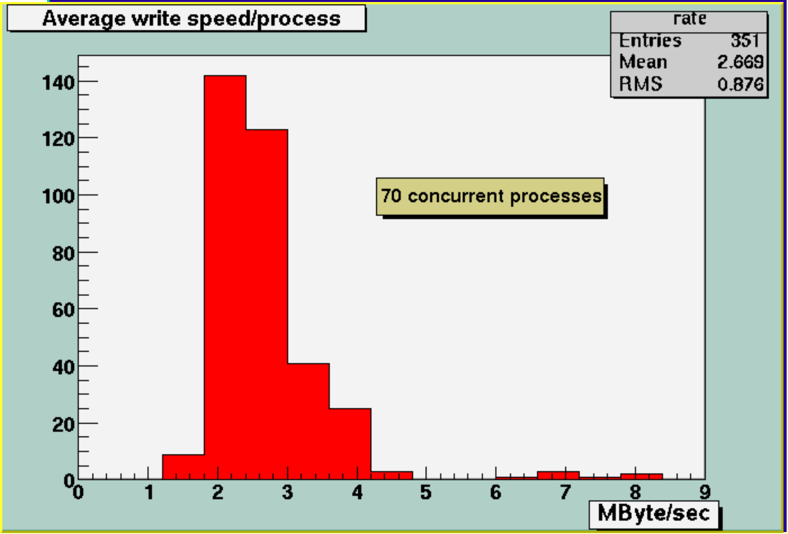
First results with dCache system

- These tests were done before the hardware and configuration upgrade. The average file size is ~1 GByte the reads are equally distributed over all read pools. Reads with dccp from popcrn nodes into /dev/null

# of concurrent reads (40 farm nodes)	Aggregate input speed (sustained over hours)
70	108 Mbyte/sec
60	104 Mbyte/sec
5	42.5 Mbyte/sec

READS

2.7 MB/sec per process

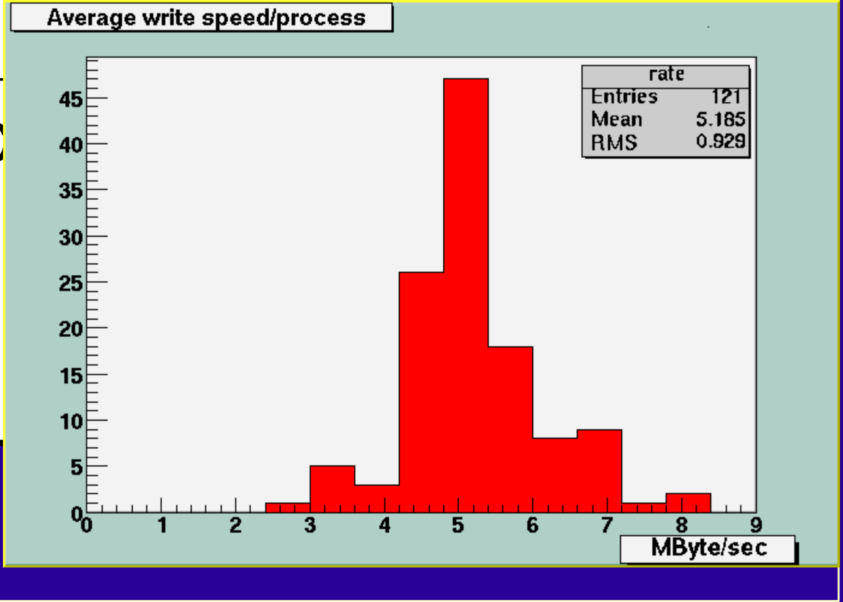


First results with dCache system

The following was done with 2 write pools. Using Root an Root Application utilizing TDCacheFile to write an Event tree into dCache. Only had three farm nodes available so probably the system is not saturated. Next test with more processes and more write pools.

# of concurrent writes (3 farm nodes)	Aggregate output speed (sustained over hours)
6	29.3 MByte/sec 5.2 MB/sec process

WRITES



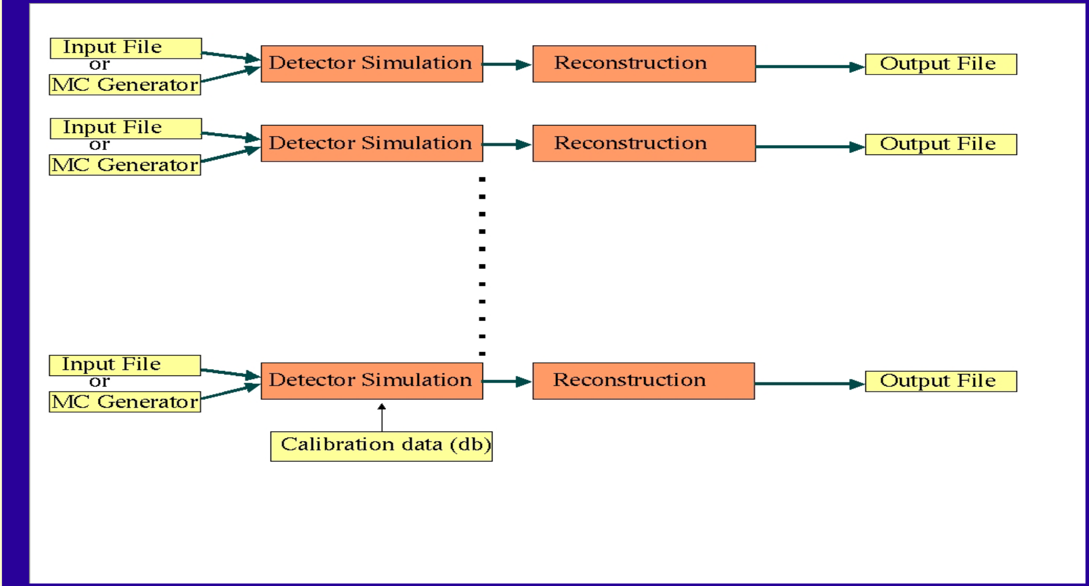
Plans for the near term future (could come up with infinite list)

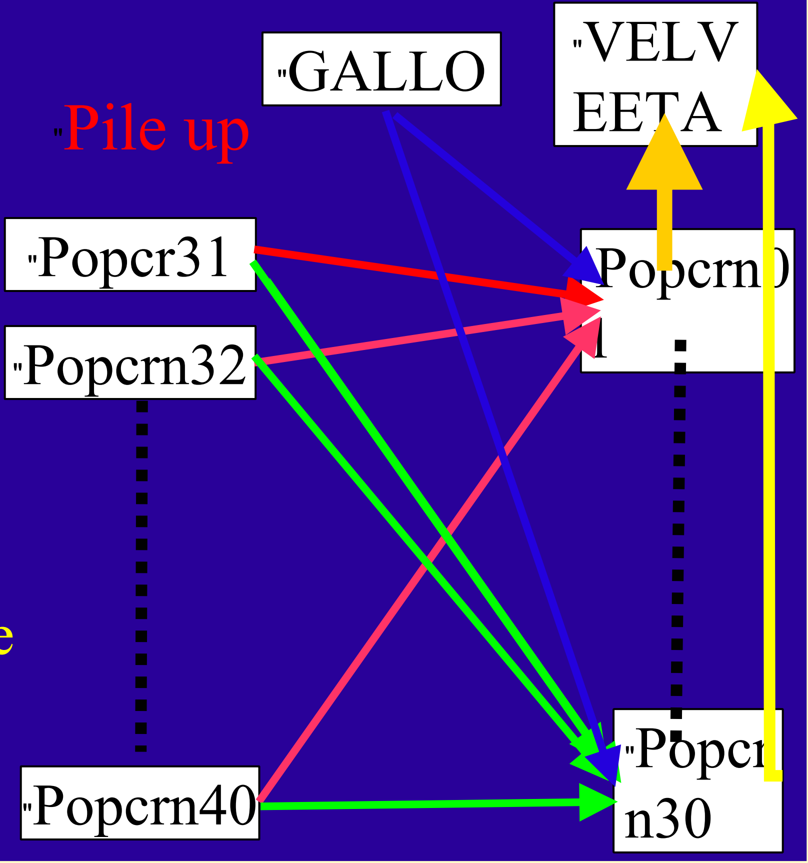
- ❑ During burn in: figure out how to configure farm for interactive use (lxplus like)(two candidates working on test farm: FBSNG, LVS). Release to users before the end of the year?
- ❑ make the user batch system easy to use. Upgrade capacity.
- ❑ Evaluate ROCKS for farm configuration.
- ❑ Allow for dynamic partitioning of farm (e.g. interactive, batch mode
- ❑ Test dCache with new EDM (rootified COBRA)
- ❑ Evaluate disk systems (zambeel etc.)

Plans for the near term future (could come up with infinite list)

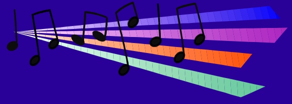
- Upgrade dCache, add new next generation servers to system
- Update monitoring
- Evaluate process monitoring tools
- …………

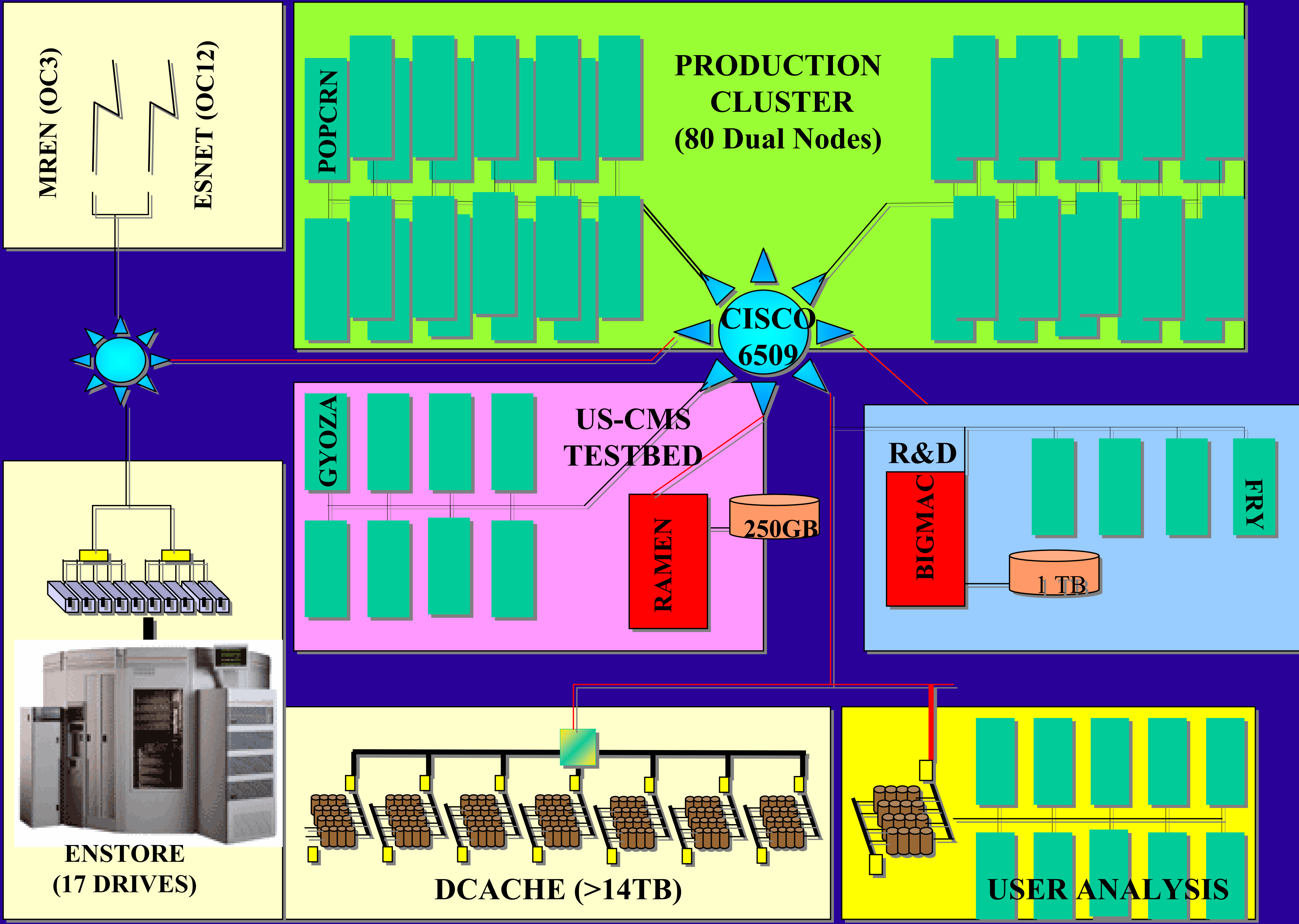
KISS





Every one talks to everyone




MREN (OC3)
ESNET (OC12)
POPCRN
PRODUCTION CLUSTER (80 Dual Nodes)
CISCO 6509
GYOZA
US-CMS TESTBED
RAMEN
250GB
R&D
BIGMAC
FRY
1 TB
ENSTORE (17 DRIVES)
DCACHE (>14TB)
USER ANALYSIS

The new farm(a single node)