Umesh Joshi
May 18, 1999

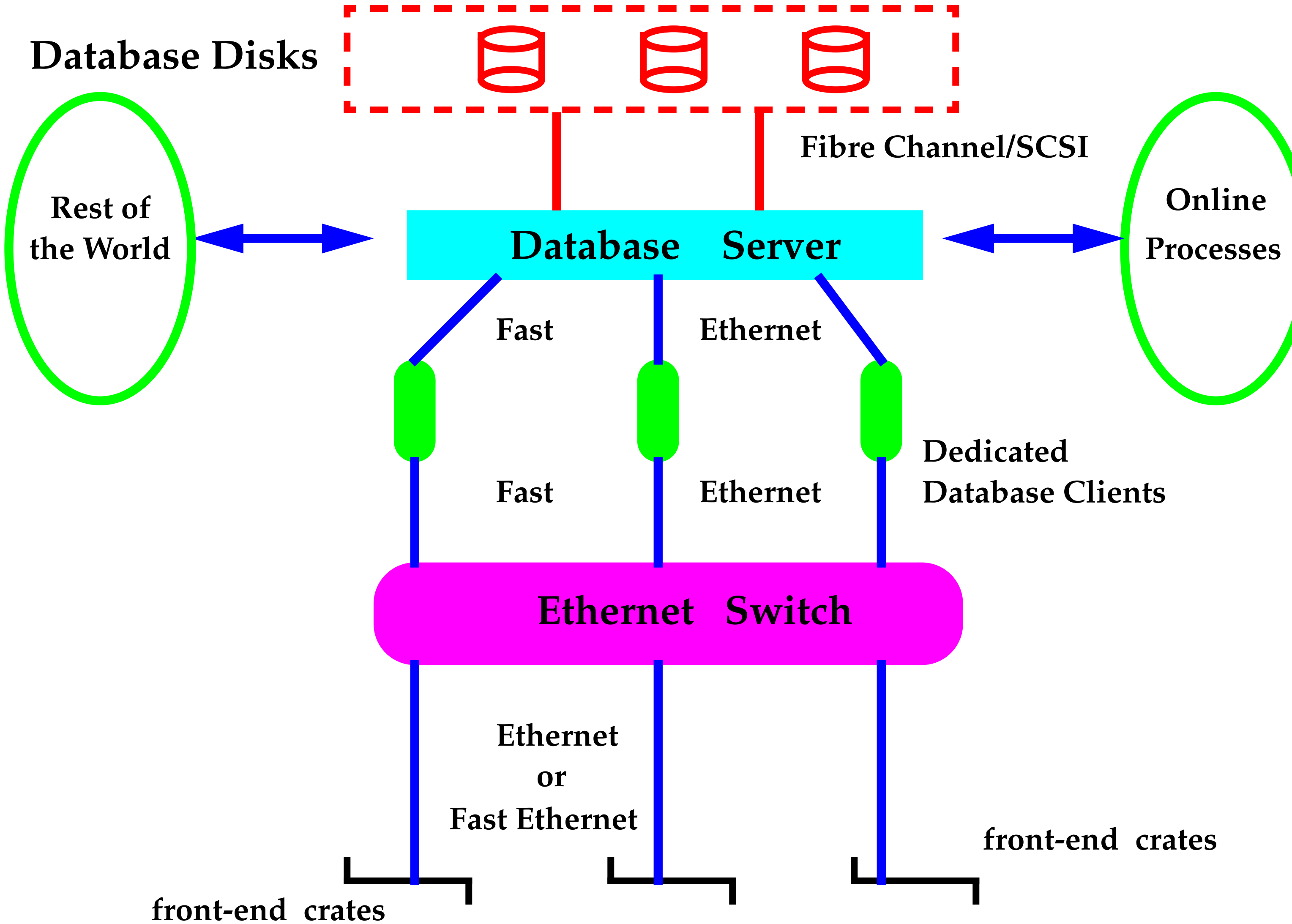
Oracle Database Studies
for
the CDF Online Database

What has been going on ?

- Work on CDF's online database in progress
- CDF's online database will be Oracle based
- Group working on APIs to run atop Oracle's existing APIs
- System model

  client-server model for data transfer between the database and front-end systems

Umesh Joshi



Logical Layout of the CDF Online Database

Studies Driven by two factors

- Performance
- System Reliability (i.e. high availability)

Performance

Desired Features

- Minimize time to start a run
- Minimize deadtime contributions.
- Rapid download of all necessary constants, programs, tables, etc.
- Concurrent calibration of different systems
- Scalability to accommodate larger amounts of data

Performance

Desired Features (Realistic Scenario)

- Actual requirements will be a moving target for a long time
- Many unknowns in procedures
- A long time before stability sets in
- System with real-time constraints: prepare for the worst case
- Database requirements are I/O and availability
- Relational aspects minimal

Performance

- Performance fundamentally limited by hardware
- Database deals with data stored on disks
- Goals of any database

  (1) optimize disk I/O
  maximize throughput
  minimize overheads

  (2) optimize network I/O (client server)
  maximize throughput
  minimize overheads

Performance

Test Environment

Two systems used to study relative performances of Oracle database

Origin 200 from **Silicon Graphics**

NUMA-Q 2000 from **Sequent Computers**

Performance

System Configuration: Origin 200 (Silicon Graphics)

Node name: ***b0dau30.fnal.gov***
4 225 MHZ IP27 Processors
CPU: MIPS R10000 Processor Chip Revision: 3.4
FPU: MIPS R10010 Floating Point Chip Revision: 0.0
Main memory size: 768 Mbytes
Instruction cache size: 32 Kbytes
Data cache size: 32 Kbytes
Secondary unified instruction/data cache size: 2 Mbytes
Integral SCSI controllers
Fast Ethernet ports

Oracle database tables all resided on a single SCSI disk

Umesh Joshi
May 18, 1999

Performance

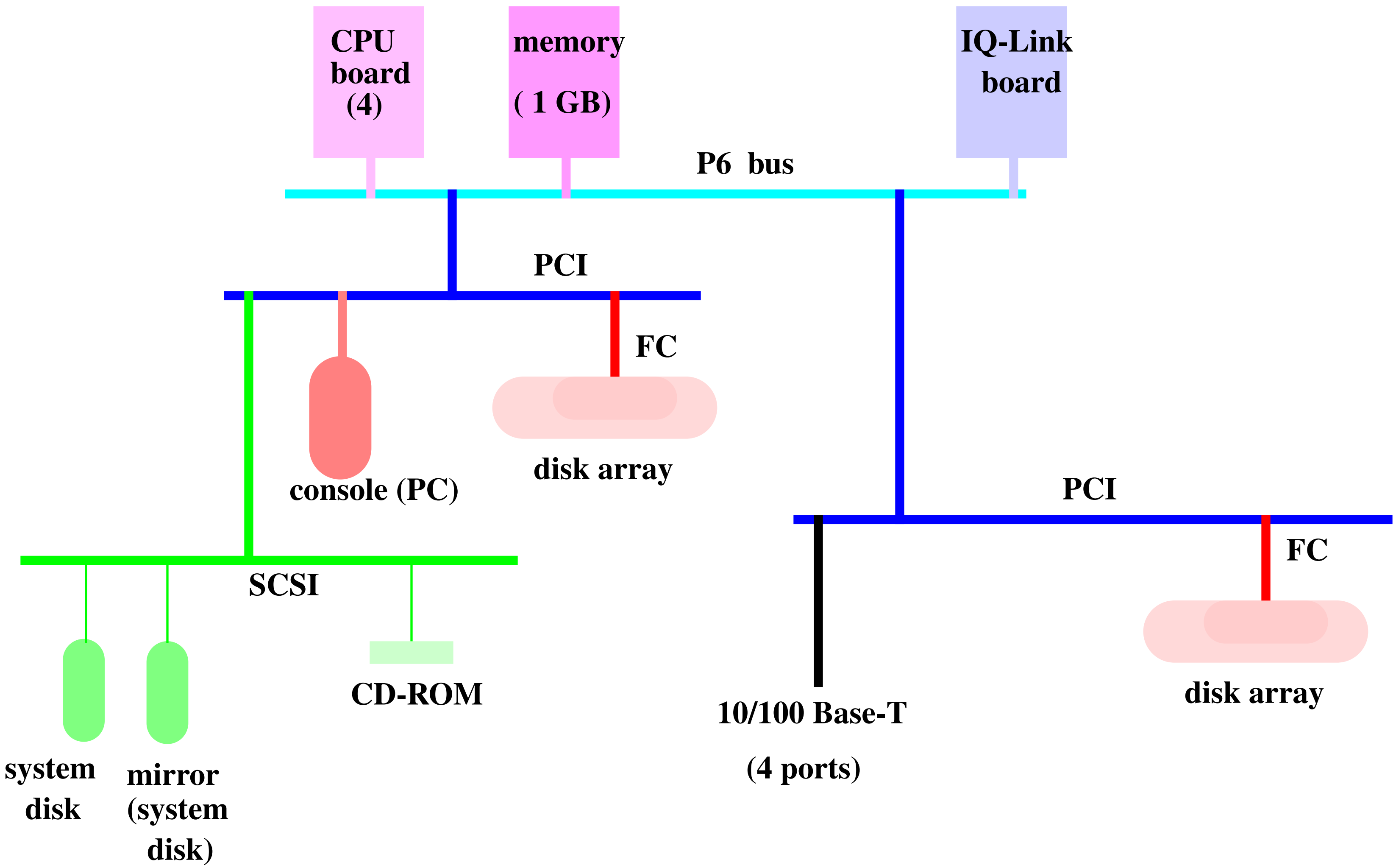
System Configuration: NUMA-Q 2000 (Sequent)

Node name: ***b0eb16.fnal.gov***
4 400 MHZ Intel Pentium II Xeon Processors
Main memory size: 1 GB
Primary (on-chip)cache size: 32 Kbytes
Secondary cache size: 1 Mbyte
System Bus: P6 bus @ 90 MHz capable of 720 MB/sec
2 64-bit PCI bus controllers @ 266 MB/sec each
1 SCSI Controller (uses 1 PCI slot)
2 Fibre Channel Controllers (use 2 PCI slots)
1 Ethernet (10/100 Base-T) Controller (uses 1 PCI slot) with 4 ports
2 CLARiiON Fibre Channel disk array enclosures (DAE) (1 per FC controller and 10 disks per array)
3 SCSI disks - system, system's mirror, and user

Performance

System Configuration: NUMA-Q 2000 (Sequent)

- Database tables can be striped across any combination of disks in the 2 CLARiiON disk arrays
- A plex can be created on the fly (Sequent Volume Manager)



Logical Layout of a NUMA-Q 2000 Quad

Umesh Joshi
May 18, 1999

Performance

CLARiiON Fibre Channel Disk Array Enclosure (DAE)

- System contains 2 DAEs (one per FC Controller)
- 10 disks per array (9.1 GB and 7200 rpm each)
- Array uses ***Fibre Channel Arbitrated Loop (FC-AL)*** over copper.

- Software data striping across multiple disks (***Sequent Volume Manager***)
- Increased I/O bandwidth
- Software striping provides great ***flexibility***, low overhead

- Number of disks in a stripe variable
- ***Allows database to be configured to different types of data requirements***, e.g. detectors with little data may not require any striping while others may require striping across a few disks.
- ***Mirroring of data*** provides ***redundancy***. Protection against disk failures.

Performance

Sequent NUMA-Q2000

- Scalable modular system
- Max of 16 modules (quads), 4 CPUs per quad
- Each module has its own I/O channels
- One instance of Oracle Database running (High availability)
- For a system with > 99% availability, need minimum of 2 quads
- All tests conducted with a single quad

Performance

Oracle Server

b0dau30.fnal.gov (SGI Origin 200)
or
b0eb16.fnal.gov (Sequent Numa-Q 2000)

Oracle Client

b0eb16.fnal.gov (Sequent Numa-Q 2000)
or
b0dau30.fnal.gov (SGI Origin 200)
or
cdfsga.fnal.gov (SGI Challenge XL)

Performance

Operating Systems

b0eb16 - DYNIX/ptx 4.4
b0dau30 - IRIX 6.5
cdfsga - IRIX 6.2

Oracle

b0eb16 - Oracle 8.0.5
b0dau30 - Oracle 8.0.4
cdfsga - Oracle 8.0.4

Performance

- Oracle Databases in *b0eb16 (Sequent)* and *b0dau30 (SGI)*

- Databases populated with CDF data from Run IB
- Slawek's files in the /data17/bottom/psia_7b/ . . . area in *cdfsga*

- Most test software in standard ***C***
- Used *Oracle Call Interface (OCI)* API

- Software to access CDF data in ***FORTRAN***
- Used **YBOS** I/O routines

Performance

- Client and server in the same machine
  b0eb16 (Sequent) or *b0dau30 (SGI)*
- Measurement of maximum I/O possible with Oracle
  (for given system configuration)

- Client and server in different systems
- Client in *b0dau30* and server in *b0eb16*
- Measurement of maximum I/O possible with Oracle
  (over a 100Base-T Ethernet for given system)

Performance

Studies

- Two tables created: TEST_YBOS4 and TEST_YBOS5
- Populated using data from Run 1B

- **TEST_YBOS4**

  each row contains data from 1 event
  16886 rows
  15 columns per row
  last column contains pointer to a location containing the whole event stored as a BLOB
  BLOB for each event (event size) ~12 KB (kilobytes)
  ~ 227 MB (megabytes) equivalent of BLOB data

Umesh Joshi
May 18, 1999

TEST_YBOS4 Schema

attribute name	datatype	description
run_num	int	run number
event_num	int	event number
lob_len	int	length of the BLOB
run_type	int	run type
part_num	int	partition number
expt_type	int	experiment type
rec_num	int	record number
rec_type	int	record type
comp_id	int	computer ID
orig_file	varchar2(30)	event type
create_date	varchar2(20)	date event recorded
create_time	varchar2(20)	name of owner
off_vers	float	offline version number
ac_vers	float	AC version number
event	blob	entire event data

Performance

- **TEST_YBOS5**

Each row contains data from 1 run
51 rows
2 columns per row
Second column contains the BLOB locator
BLOB for each run (file size) between ~5 and 10 MB (megabytes)
~ 347 MB of BLOB data in Sequent
~ 144 MB in SGI.

TEST_YBOS5 Schema

attribute name	datatype	description
run_num	int	run number
data	BLOB	entire run data

Performance

I/O Throughput

- Multiple processes: copies of tables, e.g. TEST_YBOS5_B, TEST_YBOS5_C, etc., made
- Each process accessed a different table

- Multiple disks: tablespaces created in volumes containing different number of disk stripes
- Copies of tables made in these tablespaces

- Sequent configuration allowed I/O studies with a variable number of disk stripes to be conducted.

Performance

What Kind of Performance to Expect ?

Use *dd* Unix command

Sequent

1 FC controller

1 disk: ~ 14.8 MB/sec (disk ~100% busy)
3 disk stripes : ~ 44 MB/sec (disk ~45% busy)
5 disk stripes : ~ 60 MB/sec (disk ~33% busy)
7 disk stripes : ~ 60 MB/sec (disk ~14% busy)

2 FC controllers

5 disk stripes per controller: ~ 112 MB/sec

Performance

What Kind of Performance to Expect ?

Use *dd* Unix command

Silicon Graphics

1 disk: ~ 14.5 MB/sec

Performance

Single Disk Studies (No Striping)

TEST_YBOS4: Data stored as rows and columns

Sequent

- Time to retrieve 16,886 rows (15 attributes per row): 0.63 secs
- Time to write 16,886 rows (15 attributes per row): 0.85 secs

Silicon Graphics

- Time to retrieve 16,886 rows (15 attributes per row): 1.19 secs
- Time to write 16,886 rows (15 attributes per row): 2.29 secs

Performance

Single Disk Studies (No Striping)

TEST_YBOS4: **Sequent**

- Data stored as BLOBs (~ 12 KB per BLOB)
- Total BLOB data ~ 227 MB
- Multiple disks used for multiple processes

Umesh Joshi
May 18, 1999

Performance

Single Disk Studies (No Striping)

TEST_YBOS4: **Sequent**

Number of Processes (disk id)	Data Xfer Time (secs)	**Average Xfer Rate (MB/sec)**	Number of Active FC Controllers
1 (sd17)	27.15	**8.3**	1
2 (sd17,sd1)	29.28 28.99	**15.5**	2
2 (sd17,sd18)	29.81 29.82	**15.14**	1
3 (sd18, sd1, sd18)	34.43 34.16 34.49	**19.7**	2
4 (sd18, sd1, sd2, sd17)	41.28 40.95 41.25 41.0	**22**	2

Umesh Joshi
May 18, 1999

Performance

Single Disk Studies (No Striping)

TEST_YBOS4: **Sequent**

Number of Procs	Disk Related				System Related		
	%busy	avque	avwait (msec)	avserv (msec)	%usr	%sys	**%idle**
1	38%		0	0.6	10%	12%	**78%**
2 (diff FC)	35% 37%		0	0.6 0.6	21%	28%	**51%**
2 (same FC)	38% 39%		0	0.7 0.7	21%	24%	**55%**
3	30% 34% 33%		0	0.6 0.7 0.7	34%	37%	**29%**
4	26% 31% 23% 26%		0	0.6 0.8 0.6 0.6	44%	45%	**11%**

Umesh Joshi
May 18, 1999

Performance

Single Disk Studies (No Striping)

TEST_YBOS4: Silicon Graphics

- Total BLOB data ~ 227 MB, single disk used for multiple processes

Number of Procs	Data Xfer Time (secs)	**Average Xfer Rate (MB/sec)**	Number of Active SCSI Controllers
1	40.43	**5.6**	1
2	51.6	**8.7**	1

Number of Procs	Disk Related				System Related		
	%busy	avque	avwait (msec)	avserv (msec)	%usr	%sys	**%idle**
1	70	2.1	5.9	7.0	13	33	**55**
2	44	1.8	2.9	4.6	23	42	**36**

Performance

Single Disk Studies (No Striping)

TEST_YBOS5: **Sequent**

- Data stored as BLOBs (~5 to ~10 MB per BLOB)
- Total BLOB data ~ 347 MB

Umesh Joshi
May 18, 1999

Performance

Single Disk Studies (No Striping)

TEST_YBOS5: **Sequent**

Number of Processes (disk id)	Data Xfer Time (secs)	**Average Xfer Rate (MB/sec)**	Number of Active FC Controllers
1 (sd18)	23.33	**14.9**	1
2 (sd18,sd1)	23.26 23.23	**29.8**	2
3 (sd18, sd1, sd2)	29.19 27.19 29.63	**36.3**	2
4 (sd18, sd1, sd2, sd17)	34.91 35.18 35.62 35.19	**39.4**	2

Umesh Joshi
May 18, 1999

Performance

Single Disk Studies (No Striping)

TEST_YBOS5: **Sequent**

Number of Procs	Disk Related				System Related		
	%busy	avque	avwait (msec)	avserv (msec)	%usr	%sys	%idle
1	98%	3.3	19	8.5	14%	18%	**68%**
2	98% 97%	3.0 3.0	16.7 16.8	8.4 8.4	34%	52%	**14%**
3	61% 48% 45%	1.6 1.4 1.4	4.2 2.4 1.5	6.2 5.3 5.2	46%	54%	**0%**
4	24% 30% 26% 30%	1.2 1.3 1.4 1.4	0.5 1.1 1.6 1.6	3.2 3.9 3.7 4.0	49%	50%	**0%**

Umesh Joshi
May 18, 1999

Performance

Single Disk Studies (No Striping)

TEST_YBOS5: **Silicon Graphics**

- Data stored as BLOBs (~5 to ~10 MB per BLOB)
- Total BLOB data ~ 144 MB

Number of Processes	Data Xfer Time (secs)	**Average Xfer Rate (MB/sec)**	Number of Active SCSI Controllers
1	22.83	**6.33**	1
2	76.47	**3.8**	1

Number of Procs	Disk Related				System Related		
	%busy	avque	avwait (msec)	avserv (msec)	%usr	%sys	**%idle**
1	75	1.8	5.6	7.7	13	31	**54**
2	100	3.5	24.1	12.0	21	80	**0**

Performance

Striping Across 3 Disks

TEST_YBOS5: **Sequent**

1 process: 21.2 MB/sec (%idle ~54%)
2 processes (diff FC controllers): 31.5 MB/sec (%idle ~ 6%)
3 processes: 33.2 MB/sec (%idle ~ 2%)
4 processes: 36.8 MB/sec (%idle ~0%)

3 disk stripes and 2 plexes, 3 disks per plex round robin access algorithm

TEST_YBOS5: **Sequent**

1 process: 22.5 MB/sec (%idle ~54%)
2 processes (diff FC controllers): 34.2 MB/sec (%idle ~ 6%)
3 processes: 33 MB/sec (%idle ~ 13%)
4 processes: 36.8 MB/sec (%idle ~13%)

Performance

5 disk striping

TEST_YBOS5: **Sequent**

1 process: 21.7 MB/sec (%idle ~55%)
2 processes (diff FC controllers): 34.2 MB/sec (%idle ~ 5%)
3 processes: 36.2 MB/sec (%idle ~ 2%)

10 Disk Striping

TEST_YBOS5: **Sequent**

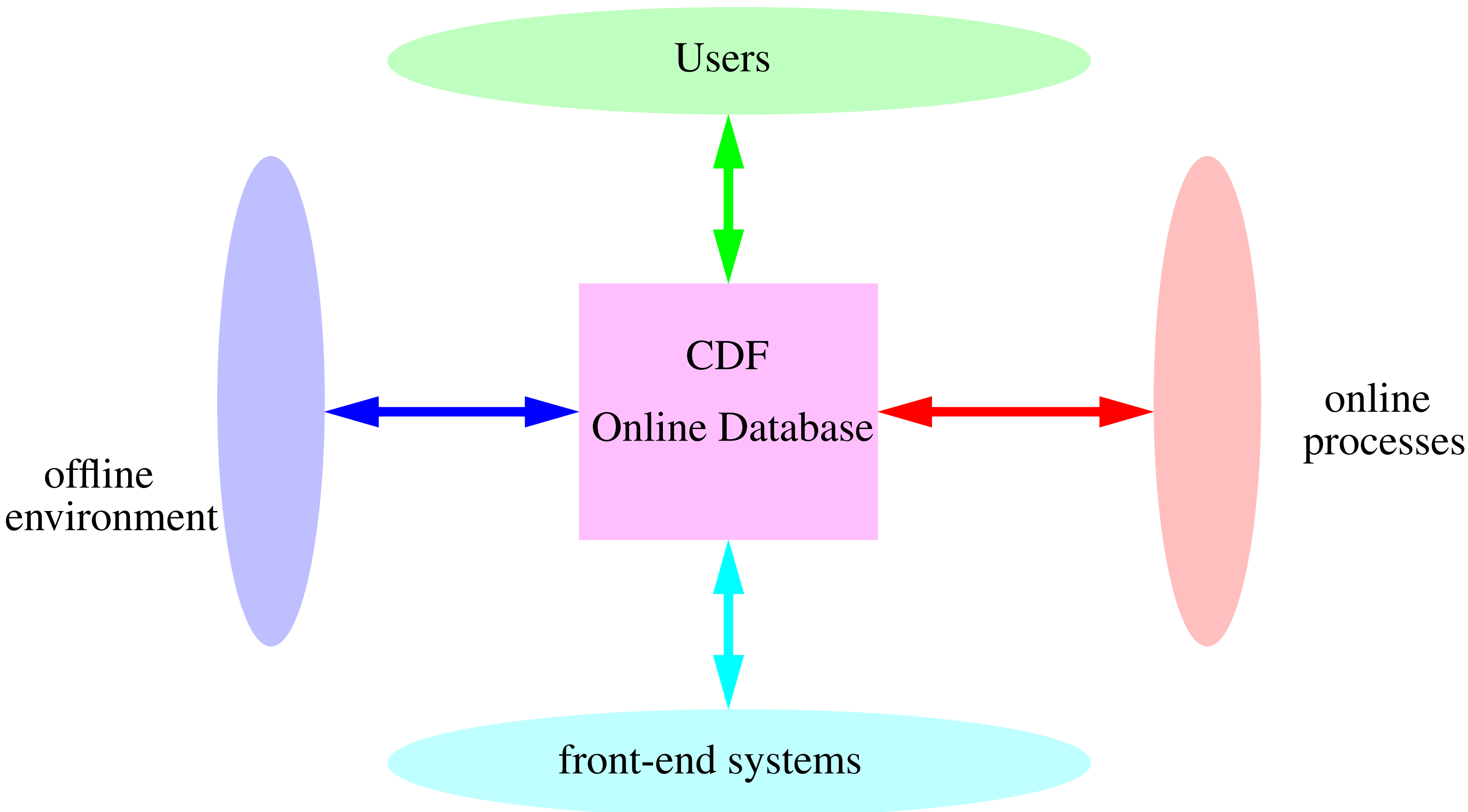
1 process: 19.8 MB/sec
2 processes: 33.3 MB/sec
3 processes: 39.4 MB/sec
4 processes: 41 MB/sec

High Availability

Does CDF need a highly available online database ?

- CDF's online database is a critical component
- Cannot take any data without it
- Should have high availability
- Highly available database

  Store every piece of information needed to start a run
  Constants, tables, executables, anything
  Will go long ways in confidence building

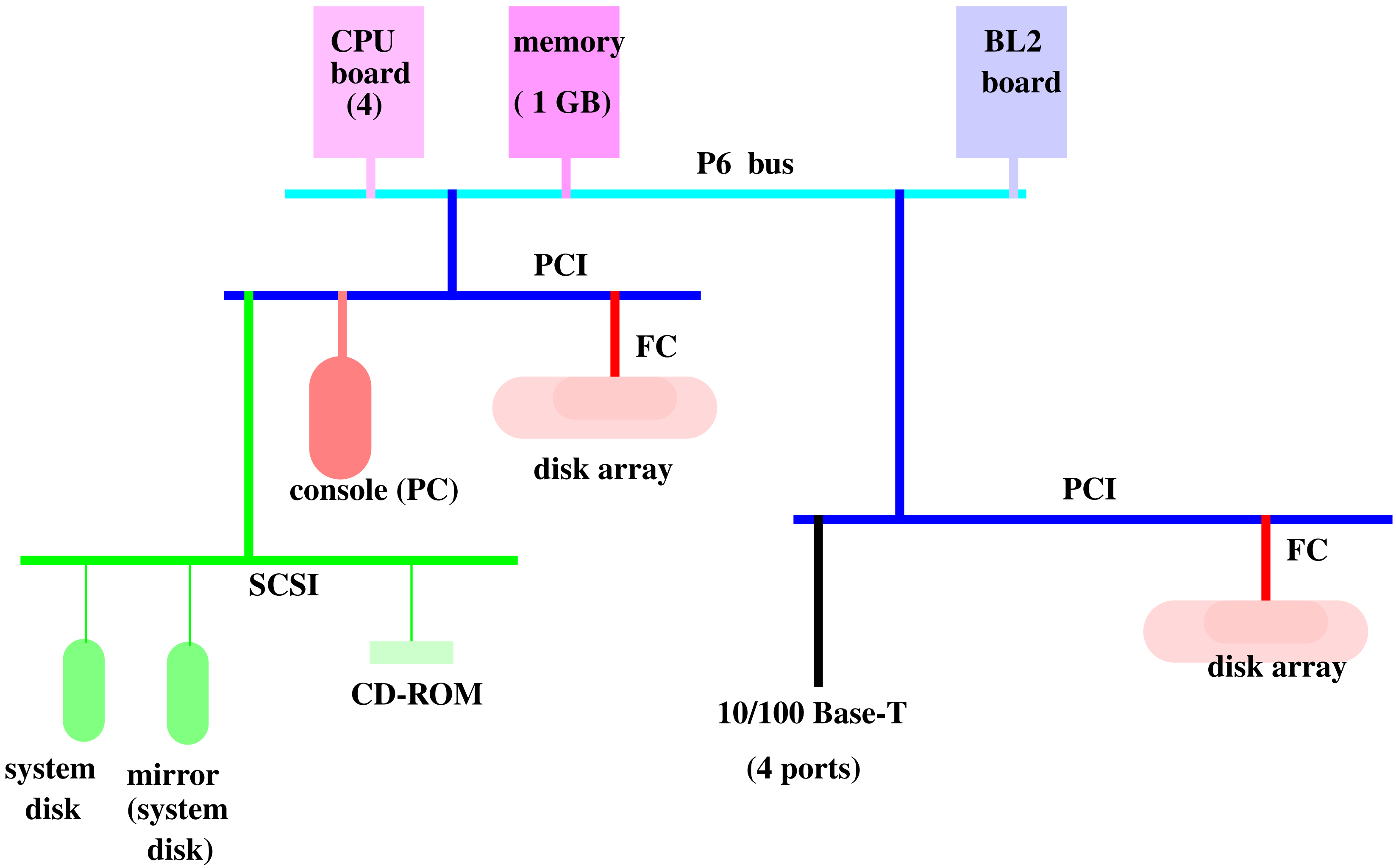


A Central CDF Database Core

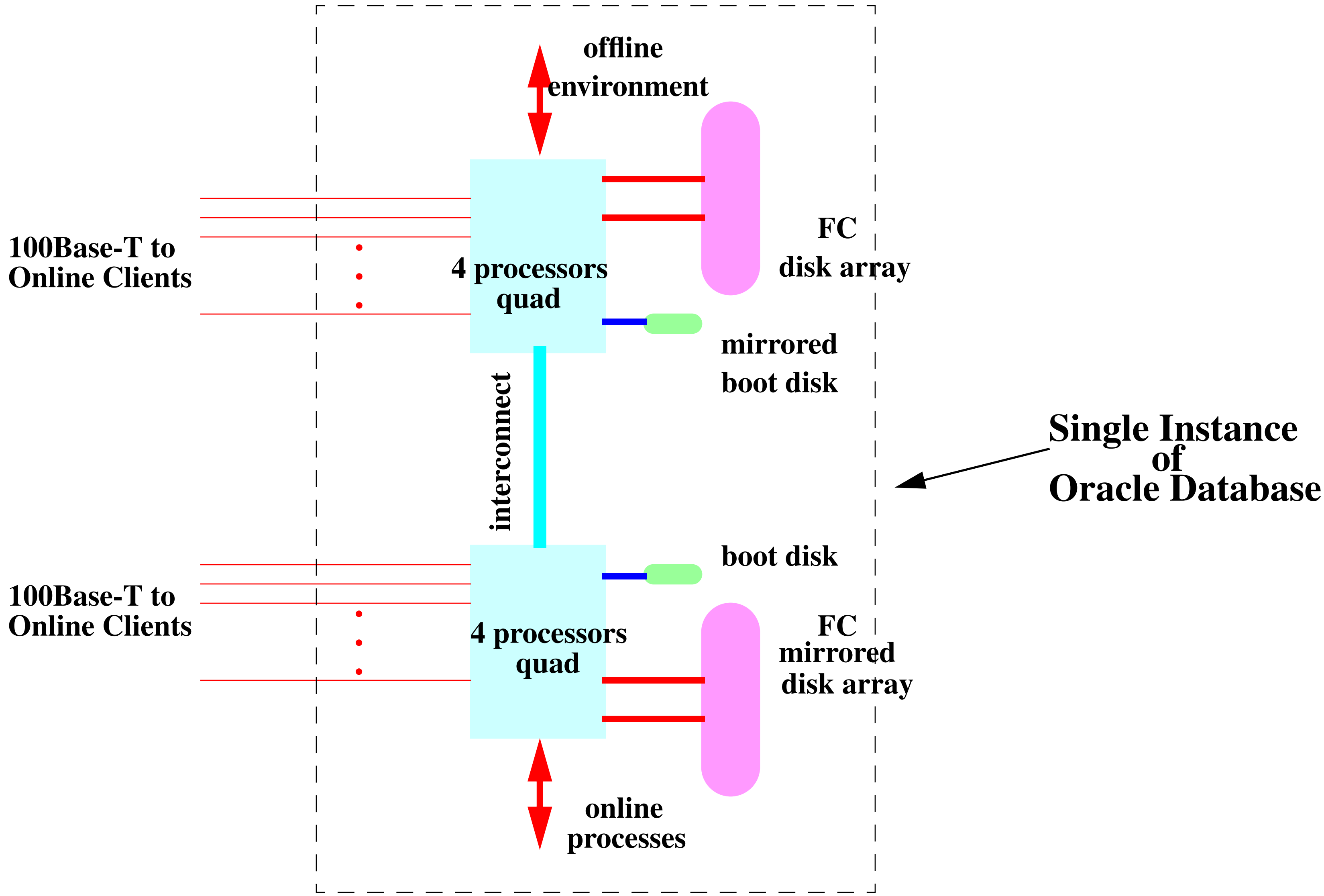
High Availability

Highly Available CDF Online Database

- Use 2 Sequent NUMA-Q 1000 quads
- Except for the interconnect, identical to NUMA-Q 2000 quad
- 500 MHz processors: higher I/O throughput
- Redundant power supplies for all modules
- Disk mirroring (system and data disks)
- Having a system with 2 quads removes single point failure
- Redundant boot system
- Redundant communications paths
- Reconfiguration around failed quad



Logical Layout of a NUMA-Q 1000 Quad


offline
environment
100Base-T to
Online Clients
4 processors
quad
FC
disk array
mirrored
boot disk
interconnect
Single Instance
of
Oracle Database
boot disk
100Base-T to
Online Clients
4 processors
quad
FC
mirrored
disk array
online
processes

High Availability

- Rolling software upgrades
- Hot database backup
- Alternate disk install
- SequentLink Service: automatically provides system health statistics and monitored automatically from the Sequent help desk.
- Console comes with a modem. For maintenance, Sequent has access to system through network and telephone line.

Features and Capabilities

2 Quad NUMA-Q 1000 System

- 500 MHz processors (8)
- Aggregate maximum Oracle I/O throughput > 80 MB/sec
- Max Oracle I/O throughput per quad > 40 MB/sec

- High availability: elimination of any single point failure
- Redundancy at every level - processing, storage, and network

- Scalability: Plenty of room for additional Ethernet and Fibre Channel cards
- 2 64-bit PCI buses per quad (4 slots per PCI bus)
- Total of 4 64-bit PCI buses (16 PCI slots in total)
- 2 independent SCSI controllers

- Limitation: maximum of 2 quads, i.e. 8 processors

Cost

Sequent NUMA-Q 1000 System

Single Quad: < $100K

- Four 500 MHz Intel Xeon Processors
- 1 GB memory
- 9 GB System Disk (SCSI)
- CLARiiON FC disk array (10 disks, each 18 GB) (PCI)
- 4 10/100 Base-T Ethernet ports (PCI)
- Console (PC with Modem)
- CDROM
- System software (Oracle not included)

Umesh Joshi
May 18, 1999

Cost

Sequent NUMA-Q 1000 System

Two Quads: < $200K

- Eight 500 MHz Intel Xeon Processors
- 2 GB memory
- Two 9 GB System disks (SCSI)
- Two CLARiiON FC disk arrays (10 disks per array, each 18 GB) (PCI)
- Eight 10/100 Base-T Ethernet ports (4 ports per card, PCI)
- Console (PC with Modem)
- CDROM
- System software (Oracle not included)