NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

Bassi
IBM POWER 5 p575

Richard Gerber

NERSC User Services Group

RAGerber@lbl.gov

June 13, NUG @ Princeton Plasma Physics Lab

About Bassi

Bassi is an IBM p575 POWER 5 cluster

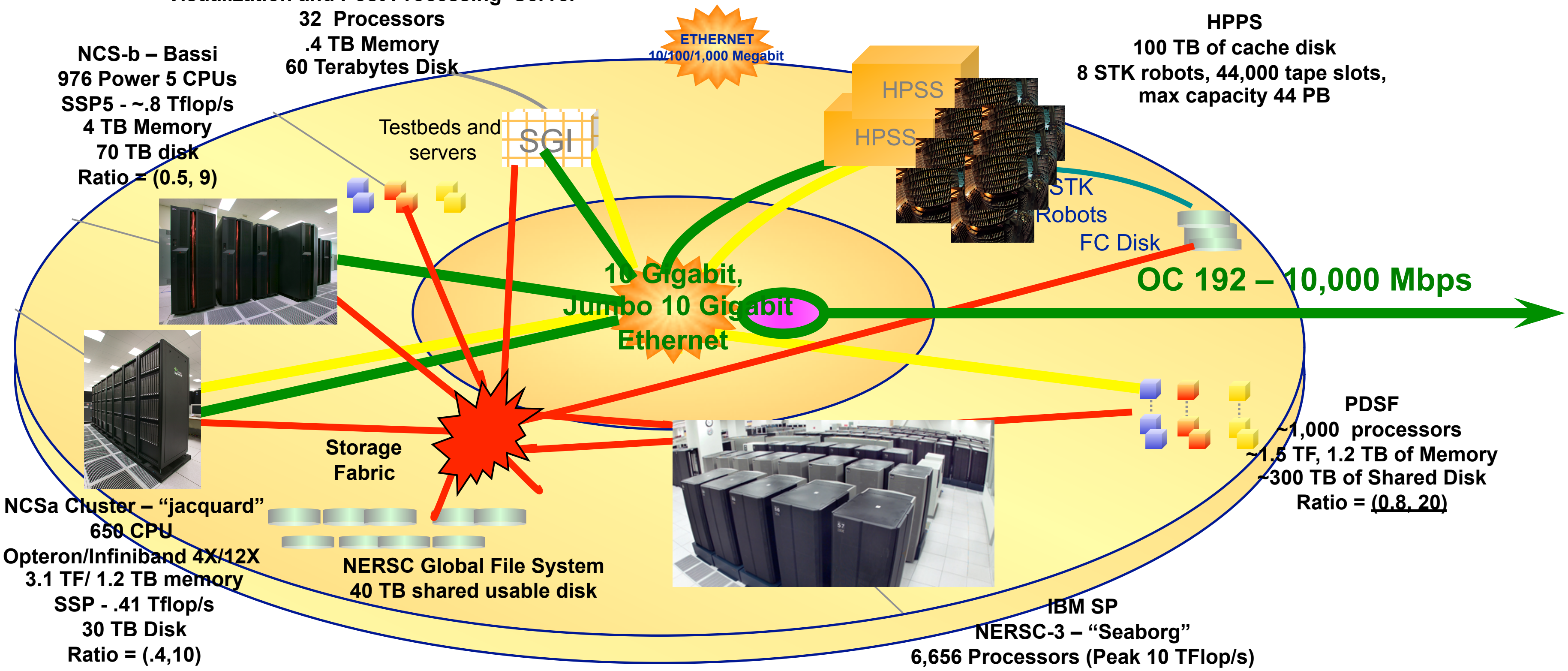
- It is a distributed memory computer, with 111 single-core 8-way SMP compute nodes.
- 888 processors are available to run scientific computing applications.
- Each node has 32 GB of memory.
- The nodes are connected by IBM's proprietary HPS network.
- It is named in honor of Laura Bassi, a noted Newtonian physicist of the eighteenth century.

Laura Bassi

Laura Bassi was perhaps the most famous woman professor at the University of Bologna. She was appointed in 1776 to the Chair of Experimental Physics. Bassi's scientific papers (one on Chemistry, 13 on Physics, 11 on Hydraulics, two on Mathematics, one on Mechanics and one on Technology), testify to the role she played in the scientific work of her age.

NERSC Configuration

January 2006



Visualization and Post Processing Server
32 Processors
.4 TB Memory
60 Terabytes Disk

NCS-b – Bassi
976 Power 5 CPUs
SSP5 - ~.8 Tflop/s
4 TB Memory
70 TB disk
Ratio = (0.5, 9)

HPPS
100 TB of cache disk
8 STK robots, 44,000 tape slots,
max capacity 44 PB

PDSF
~1,000 processors
~1.5 TF, 1.2 TB of Memory
~300 TB of Shared Disk
Ratio = (0.8, 20)

NCSa Cluster – “jacquard”
650 CPU
Opteron/Infiniband 4X/12X
3.1 TF/ 1.2 TB memory
SSP - .41 Tflop/s
30 TB Disk
Ratio = (.4,10)

NERSC Global File System
40 TB shared usable disk

IBM SP
NERSC-3 – “Seaborg”
6,656 Processors (Peak 10 TFlop/s)
SSP5 – .9 Tflop/s
7.8 Terabyte Memory
55 Terabytes of Shared Disk
Ratio = (0.8,4.8)

Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)

Bassi's Role at NERSC

- Bassi serves the needs of scientists with codes that scale somewhere between those that run on Jacquard and Seaborg.
- The target parallel concurrency is 64-256 MPI tasks.
- It is relatively easy for Seaborg users to port and run their codes, because Bassi has a familiar computing environment.

Bassi System Configuration

- 122 8-processor nodes (with 32GB memory each)
- 111 compute nodes (888 processors)
- 3.5 TB aggregate memory on compute nodes
- 7.6 GFlops/sec peak processor speed
- 6.7 TFlops theoretical peak system performance
- 100 TB of usable disk space in GPFS (General Parallel Filesystem from IBM)
- 2 login nodes
- 6 VSD (GPFS) servers
- The nodes are configured to use 24 GB of "Large Page" memory

Bassi System Specs

NERSC Bassi Nodes

IBM designation	p575
Processor (single core)	POWER 5
Processor Speed	1.9 GHz
Number of CPUs per node	8
Physical memory per node	32 GB
Number of network adapter cards for inter-node communication	1 (2-link)

Bassi System Specs

POWER 5 Processor

Clock speed	1.9 GHz
FP Results/Clock	4
Peak Performance	7.6 Gflops
L1 Instruction Cache	64 KB
L1 Data Cache	32 KB
L2 Cache	1.92 MB
L3 Cache	36 MB
Packed-Node Memory Bandwidth per CPU	7 GB/s (20X Seaborg)

Bassi Memory Configuration

- Each node has 32 GB of memory shared by the 8 CPUs.
- 24 GB is configured as “large page” memory (16 MB pages); reduces TLB misses; HPC codes run about 20% faster on average.
- Binaries must be “large-page” enabled, which is the Bassi default (but if you override the NERSC default, you’re on your own! Large page memory is not available to non-enabled binaries, so you will have only ~2 GB/node available)
- MEMORY_AFFINITY=MCM keeps memory “close” to CPU

HPS Interconnect (Federation)

- Custom IBM interconnect, named HPS (aka “Federation)
- Dual plane; separate connect to each from each node
- Latency of <4.4 μs, ~5 times better than Seaborg
- Measured point-to-point bandwidth > 3.1 GB/s unidirectional, 10 times greater than Seaborg
- Theoretical HPS bandwidth 2 GB/sec per link each direction.

Lots of Information on the Web

- Go to http://www.nersc.gov/nusers/resources/bassi/

Bassi Delivery and Acceptance

- System delivery started 7/11/2005; system was integrated on-site.
- Because of power limitations, software was installed frame by frame, with switch integration after facility power upgrade completed
- Acceptance period began 10/14/2005; system was accepted on 12/15/2005.
- System availability ended with 99% + availability and 86% + utilization.
- Bassi went into production 01/09/2006.

Bassi Authentication

- Your Bassi password is your NERSC LDAP password. This is also your NIM password. Password changes are done through the NIM web interface.
- This has caused many problems, due to incomplete (and buggy) IBM implementation.
  - Many problems with user filegroup, repo, shell information
  - A side-effect of AIX/PE problems has caused recent job launch failures.

Bassi Environment

- A full instance of AIX 5.3D is running on each node. Uses ~ 5 GB (mostly small page memory)
- 64-bit code builds are the default (OBJECT_MODE=64)
- NERSC sets many environment variables to default values that help “typical” codes.
- Two you may want to override:
  - MP_TASK_AFFINITY=MCM binds MPI tasks to CPUs, but breaks OpenMP codes (solution: unsetenv MP_TASK_AFFINITY)
  - MP_SINGLE_THREAD=yes for codes that are known to be single-threaded helps performance, but breaks the threaded MPI-IO and MPI-2 one-sided functions (unsetenv MP_SINGLE_THREAD)

Important Run-Time Settings

- https://www.nersc.gov/nusers/resources/bassi/running_jobs/architecture.php

Bassi Compilers and Libraries

- The AIX compilers should be familiar to Seaborg users.
- GCC is available, but recommended only when AIX compilers won't do (module load gcc)
- The libraries you expect are there: ESSL, NAG, Scalapack, etc.; 64-bit builds are the default, but 32-bit symbols are in there two where possible.

Running Jobs

- Parallel jobs are run under POE and LoadLeveler, just as on Seaborg.
- The submit classes are regular, low, premium, debug and interactive.
- The charge factor is 6 for regular, 3 for low and 12 for premium.
- Jobs up to 48 nodes running for 12 hours (24 hours for <16 nodes) are accommodated normally.
- Larger, longer-running jobs are allowed upon request.

Bassi Queues (Classes)

Submit Class[1]	Destination Class[2]	Nodes	Max Wallclock	Rel Priority	Availability
interactive	interactive	1-4	30 mins	1	Everyone
debug	debug	1-8	30 mins	2	Everyone
premium	premium	1-48	12 hrs	4	Everyone
regular	reg_1	1-15	24 hrs	5	Everyone
	reg_16	16-31	12 hrs	5	Everyone
	reg_32	32-48	12 hrs	5	Everyone
low	low	1-32	12 hrs	6	Everyone
special*	special	1-64	48 hrs	3	By special arrangement
full_config*	full_config	1-ALL	48 hrs	3	By special arrangement

Bassi Filesystems

- $HOME quota is 5 GB per user
- $SCRATCH quota is 250 GB per user
  - Tuned to achieve 4 GB/sec R&W aggregate bandwidth from 32 tasks (not packed).
- /project (NGF) is mounted
- HPSS available via the usual HIS and PFTP utilities
- Quotas are "group" quotas on your "personal filegroup," not user quotas. (This might be confusing if you don't realize it.)
  - "myquota" command will show your (group) quota by default, but don't use "myquota –u username"

Bassi Benchmark Suite

- The SSP for Bassi consists of 6 codes, whose performance is averaged and scaled to the system size. There are two classes of codes:
  - 3 NAS Parallel Benchmarks: a well-tested standard set of computational kernels.
  - 3 NERSC user codes
    - CAM 3: Atmospheric climate model
    - GTC: Fusion turbulence code
    - PARATEC: Material Sciences code
- Most are run using 64 MPI tasks.

SSP Results

Code	SOW Commitment (Mflops/s/task)	Measured (as delivered)	Performance Ratio vs. Seaborg
NPB FT	670	822 (673)	8.95
NPB MG	800	1345 (889)	8.86
NPB SP	480	572 (492)	9.56
CAM	493	554 (517)	4.85
GTC	650	753 (658)	5.19
PARATEC	4400	4794 (4304)	5.65

SSP Results II

- IBM proposed a .75 TFlops/sec system as measured by the SSP.
- With fixes, tuning, and configuration changes during the acceptance period, Bassi’s SSP is about .90-.92 TFlops/sec for 888 processors.
- For comparison, Seaborg, with 6,080 processors, measures .916 TFlops/sec on the Bassi SSP code suite.

Non-Dedicated Benchmark Performance

- Bassi’s performance in non-dedicated mode is similar to dedicated performance, with very small variation.

	N Trials	Time	Performance	COV	Required
NPB FT PAR	87	171.45	817.20	1.81%	670
NPB MG PAR	79	36.35	1338.58	0.54%	800
NPB SP PAR	42	777.27	593.74	0.59%	480
CAM 16x1	87	1388.56	501.12	0.17%	493
GTC	90	163.12	750.86	0.77%	650
PARATEC	88	599.10	4721.82	1.77%	4400
SSP/task			1024.15		844.51
SSP			0.909		0.750

Micro and Misc Benchmarks

CAM 16x2	92	734.83	947.10	0.66%	954
CAM 16x4	89	397.51	1750.61	0.87%	1776
FT SERIAL	106	90.71	1014.84	0.26%	886
MG SERIAL	86	13.45	1447.70	1.65%	1272
SP SERIAL	77	553.61	641.43	2.63%	638
MPI PP LATENCY	83		4.7338	1.04%	4.768
MPI PP BW	83		3099	1.43%	1186
MPI ORD RING BW	83		1842	6.89%	740
MPI RND RING BW	83		260	5.83%	172
MEMRATE SINGLE	117		7212	0.52%	5218
MEMRATE MULTI	117		6851	1.63%	5228
PIORAW READ	78		4401	5.66%	4000
PIORAW WRITE	78		3984	0.93%	4000

Bassi Status and Open Issues

- Bassi is running AIX 5.3 at AIX 5.2 performance levels (this was not easily accomplished!)
- There are still unresolved authentication issues, but we hope they are currently transparent to you and will continue to be so.
- No major problems known, but many minor problems are just now being addressed because the AIX 5.3 migration put them on the back burner.
  - SMT testing
  - UPC
  - LL/PE bugs
  - Website updates have been deferred; hope to document and track outstanding issues on Bassi pages very soon
  - Occasional MPI timeouts have been observed.
  - NGF performance testing and tuning
  - etc

Additional Information

- The web page for Bassi users is:
  - http://www.nersc.gov/nusers/resources/bassi/