Supercomputers

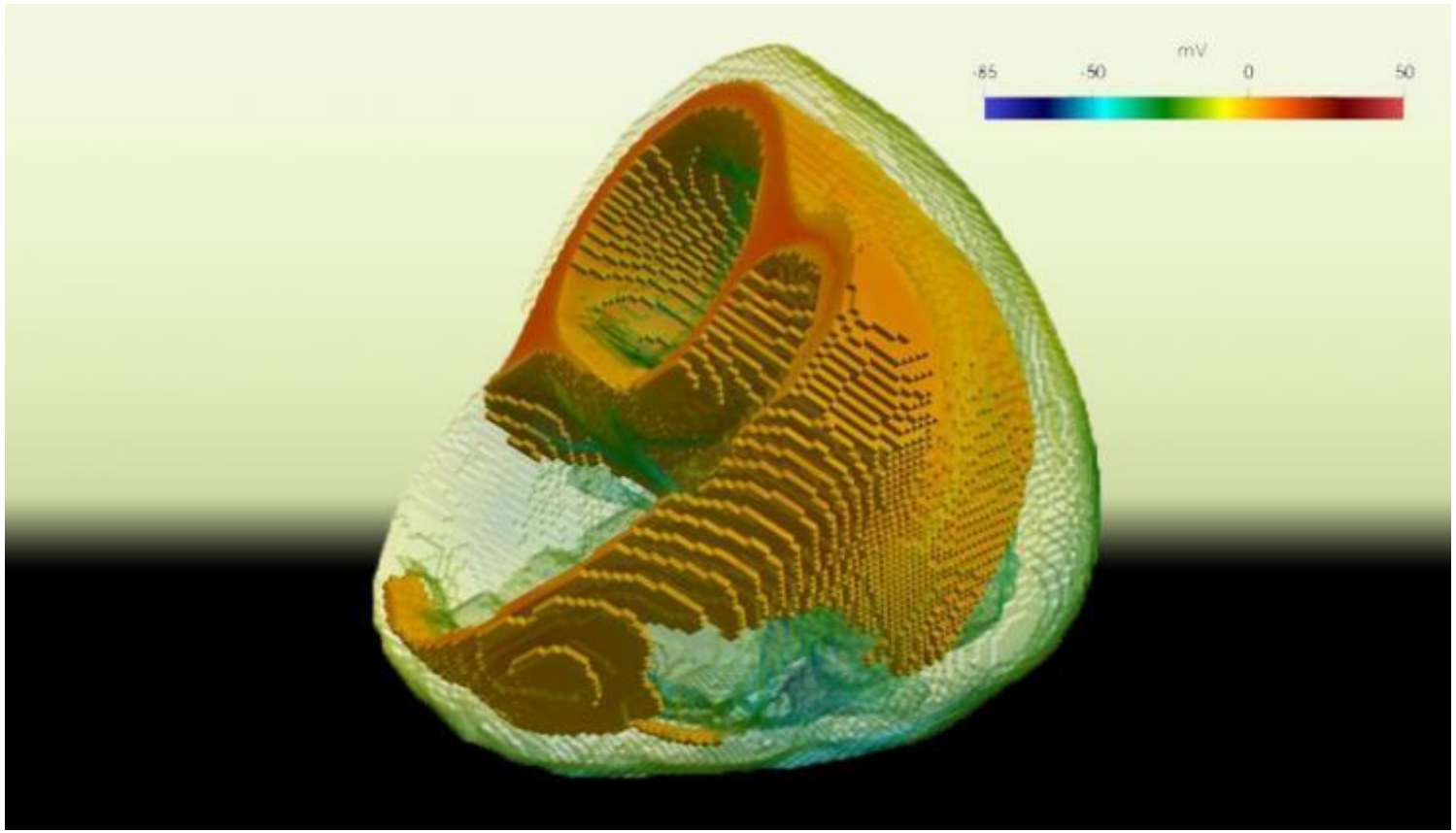
Apr 23, 2021

IBM Blue Gene/Q

- November 2011
  - 4,096-node BG/Q (Sequoia)
  - #17 on top500 at 677.10 TF
  - #1 Graph 500 at 254 Gteps (Giga traversed edges/second)
  - #1 on Green 500 list at 2.0 Gflops/W
- June 2012
  - #1 Sequoia at Lawrence Livermore National Laboratory (now #13)
    - 96K nodes, 16.3 PF Max, 20 PF Peak, 7.8 MW
  - #3 Mira at Argonne National Laboratory (now #24)
    - 48K nodes, 8.1 PF Max, 10 PF Peak, 3.9 MW
    - Decommissioned in Dec 2019

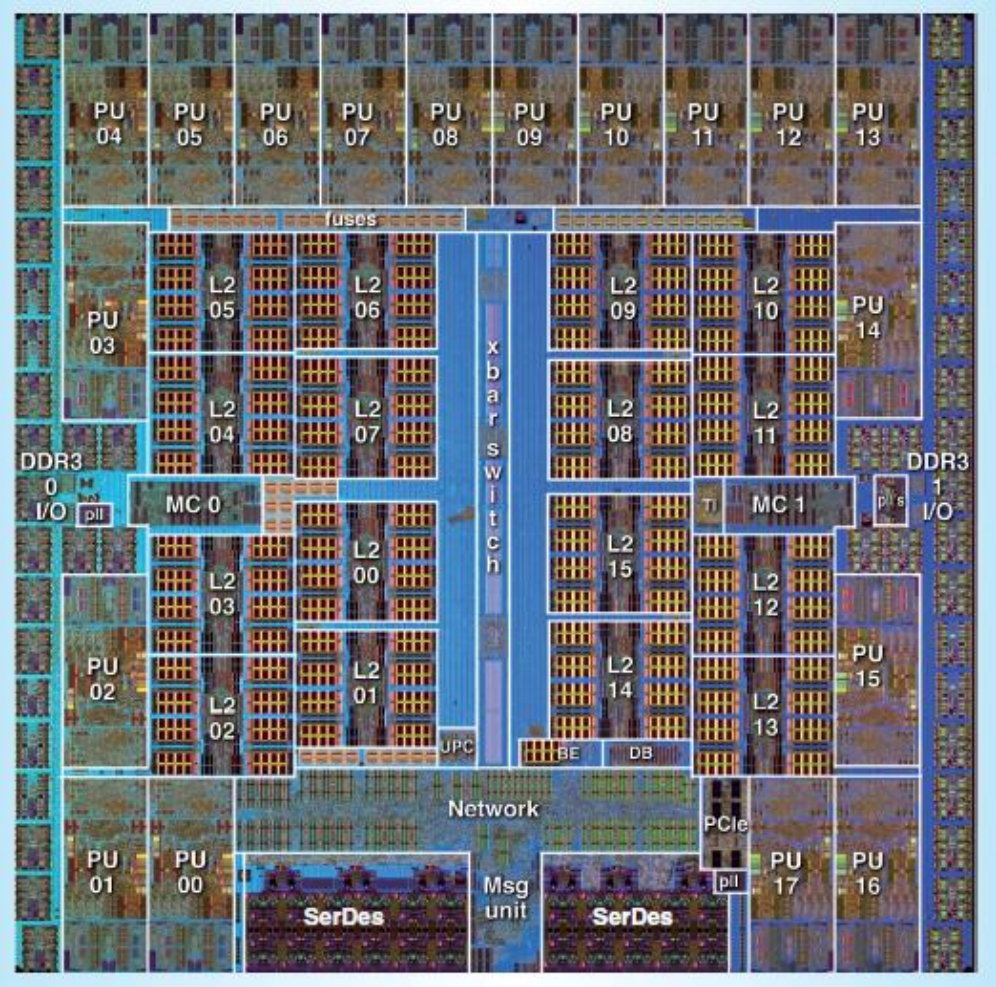
Real Applications on Sequoia

Cosmology code HACC 14 PFLOPS



Heart simulation code Cardioid 12 PFLOPS

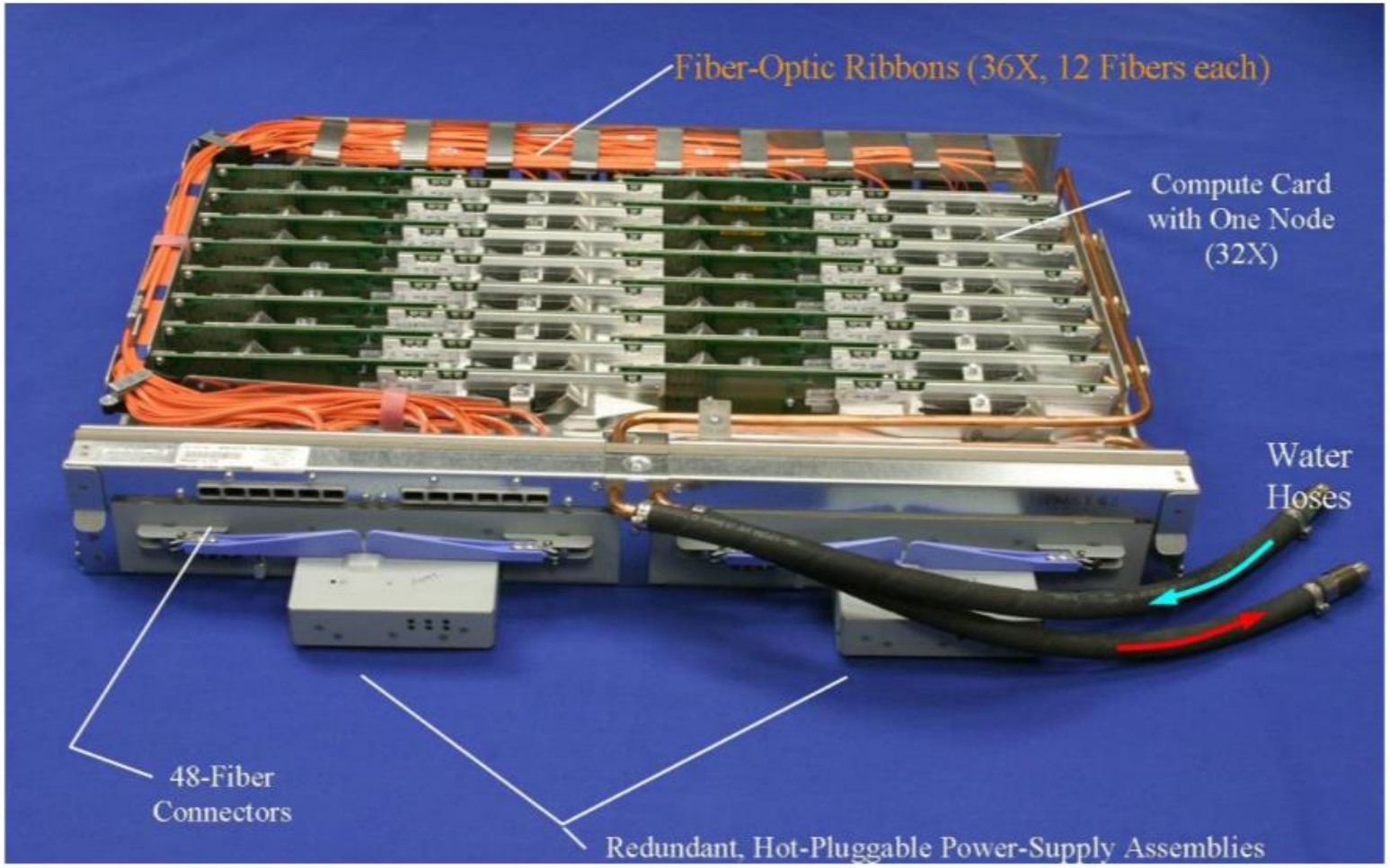
Blue Gene/Q Compute Chip (BQC)



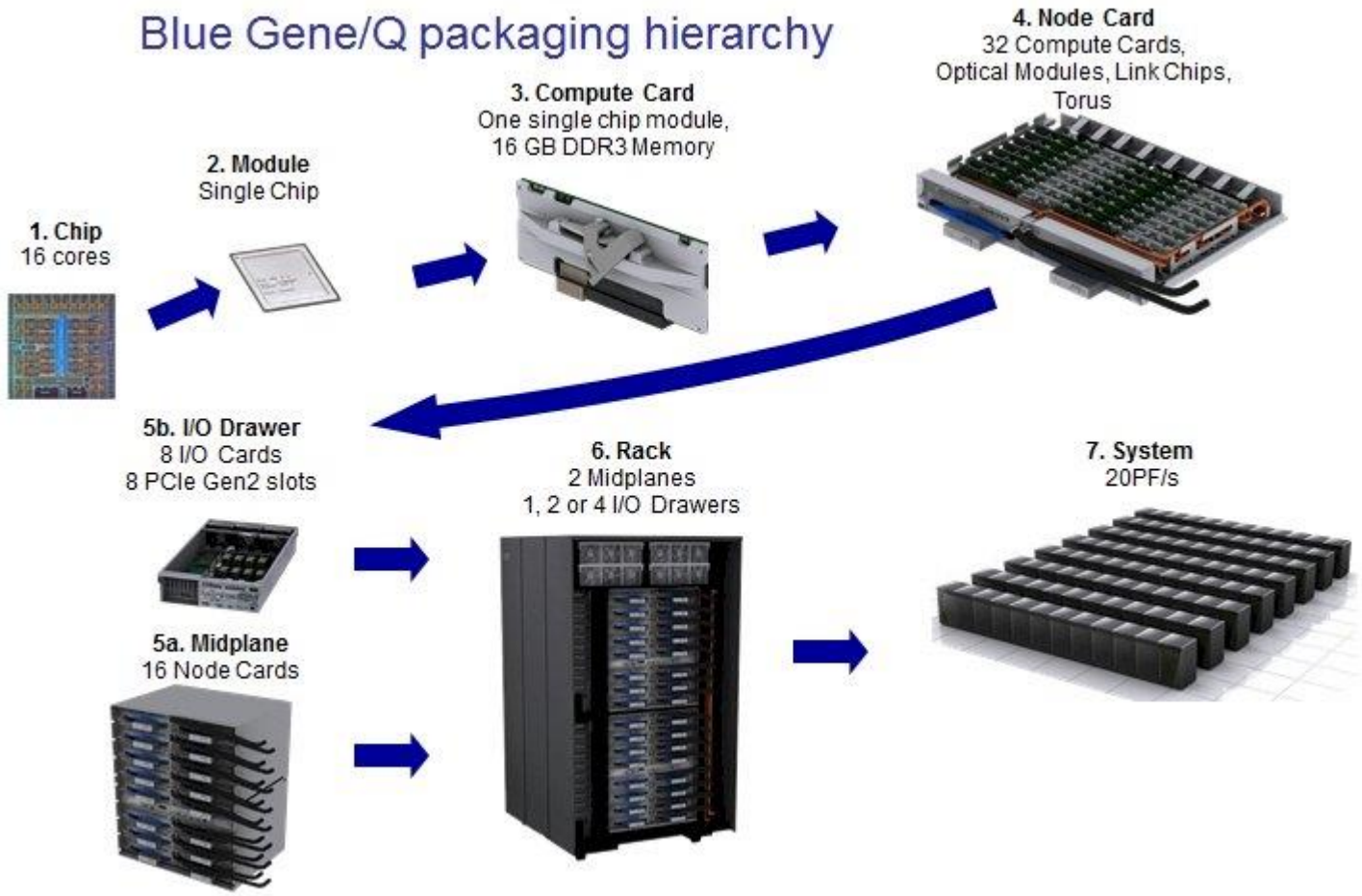
- 18.96 x 18.96 mm chip (45 nm, 1 billion transistors)
- Processors, memory and network on chip
- PowerPC A2 Processor Core
  - 1.6 GHz
  - 64-bit Power ISA
  - In order execution
  - 4-way SMT
  - 2-way concurrent instruction issue
- Quad FPU

The IBM Blue Gene/Q Compute Chip, IEEE MICRO, 2012

Compute Node Board



BG/Q Hierarchy



- Rack (1024 nodes), midplane (512 nodes), node-board (32 nodes)
- Jobs may run independently on partitions

Interconnects in BG

- BG/L has a 3D torus with 175 MB/s per link
- BG/P has a 3D torus with 425 MB/s per link
- BG/Q has a 5D torus with 2 GB/s per link

Why 5D torus?

- Lower diameter, higher bisection width, lower latency than 3D torus
- High nearest neighbour bandwidth

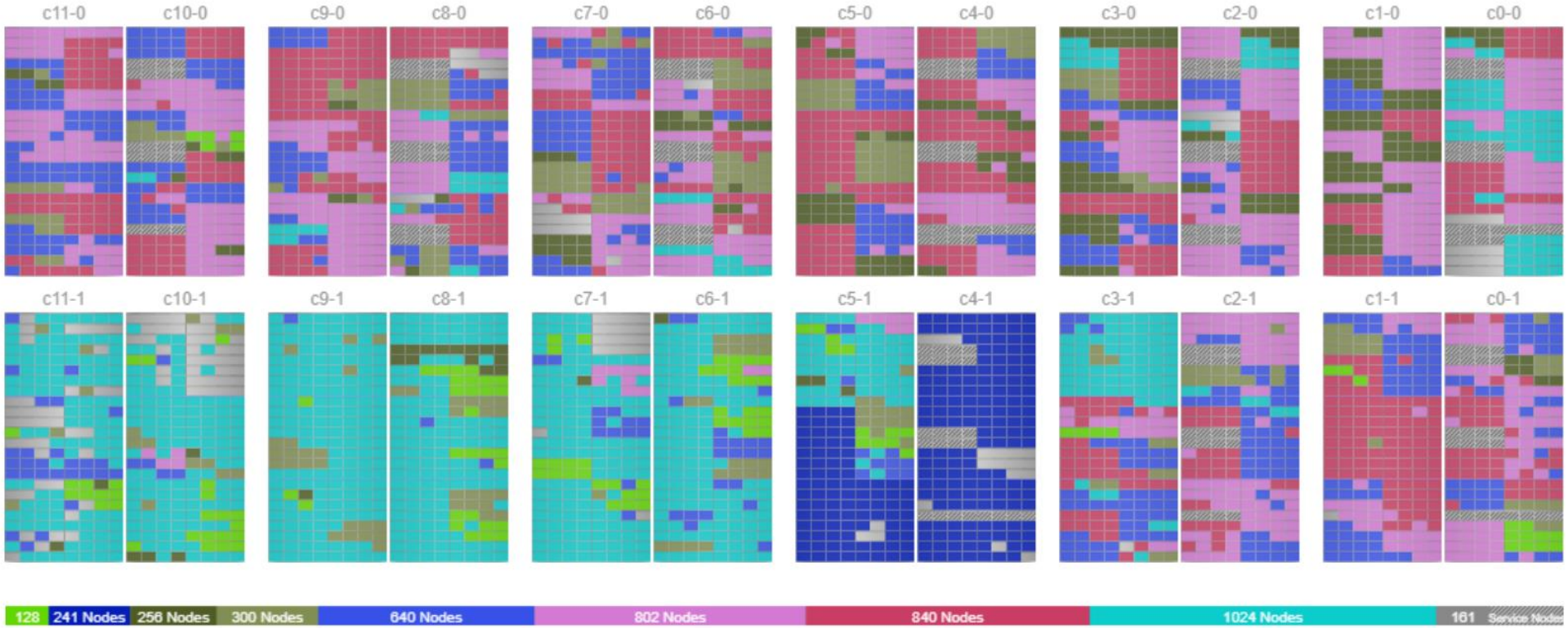
BG/Q Messaging Unit and Network Logic

- A, B, C, D, E dimensions (5D torus)
  - Last dimension E is of size 2 (reduces wiring)
  - Link chips on each node board connect via optics to node boards on other midplanes
- On-chip per hop latency: 40 ns (20 network cycles)
  - 16x16x16x12x2 P2P latency is expected to be about 2.6 µs, including cable delays
  - 0.6 µs at 1 hop, 1.17 µs at 13 hops
- Injection and reception FIFOs (More than half latency incurred here)
  - E.g. Packets arriving on A- receiver are always placed on A- reception FIFO

References for BG/Q

- The IBM Blue Gene/Q Compute Chip, IEEE MICRO, 2012.
- The IBM Blue Gene/Q Interconnection Fabric, IEEE MICRO, 2012.
- The IBM Blue Gene/Q Interconnection Network and Message Unit, SC 2011.
- Looking Under the Hood of the IBM Blue Gene/Q Network, SC 2012.
- IBM System Blue Gene Solution: Blue Gene/Q Application Development, IBM Redbooks, 2013.

Supercomputer Job Allocation



Running | Starting | Queued | Reservations

Total Running Jobs: 8

Job Id		Project	Nodes	Start Time	Run Time	Walltime	Queue	Mode
512304		EstopSim_2	1024	9:13:30 AM	00:12:13	16:00:00	default	script
512623		TurbShockWalls	840	7:57:42 AM	01:28:01	1d 00:00:00	default	script
498557		TurbShockWalls	802	9:43:57 PM	11:41:46	1d 00:00:00	default	script
513358		PSFMat_2	640	8:29:27 AM	00:56:16	12:00:00	default	script
511830		ReconDepth	300	7:50:53 AM	01:34:50	06:00:00	default	script
514000		HighLumin	256	8:50:22 AM	00:35:21	06:00:00	default	script
514114		CVD_CityCOVID	241	1:28:47 AM	07:56:56	1d 12:00:00	CVD_Research	script
514178		FDTD_Cancer_2a	128	8:00:51 AM	01:24:52	03:00:00	default	script

Status.alcf.anl.gov

Total Queued Jobs:

Job Id	Project	Score	Walltime	Queued Time	Queue	Nodes	Mode
488198	NucStructReact_6	40.00	06:00:00	114d 02:05:56	backfill	2048	script
488200	NucStructReact_6	40.00	06:00:00	114d 02:05:33	backfill	2048	script
488201	NucStructReact_6	40.00	06:00:00	114d 02:05:15	backfill	2048	script
488202	NucStructReact_6	40.00	06:00:00	114d 02:04:52	backfill	2048	script
488612	NextGenReac	40.00	06:00:00	110d 10:35:36	backfill	1024	script
505064	TurbShockWalls	86.10	09:00:00	33d 20:13:16	default	512	script
509215	ClimateEnergy_4	59,231.06	01:00:00	20d 05:04:52	default	128	script
511951	TurbShockWalls	1,074.68	1d 00:00:00	8d 02:01:15	default	802	script
512878	CSC249ADSE16	397.24	09:30:00	5d 00:47:21	default	1024	script
513149	NanoReactive_3	1,120.58	14:00:00	4d 06:39:18	default	1041	script
513267	TurbShockWalls	620.78	1d 00:00:00	3d 17:12:12	default	840	script
513422	CSC249ADCD08	15.44	06:00:00	3d 03:43:45	backfill	256	script
513426	CSC249ADCD08	15.39	06:00:00	3d 03:25:28	backfill	256	script
513434	CSC249ADCD08	15.18	06:00:00	3d 02:05:25	backfill	256	script
513435	CSC249ADCD08	15.18	06:00:00	3d 02:05:10	backfill	256	script
513436	CSC249ADCD08	15.18	06:00:00	3d 02:04:42	backfill	256	script
513437	CSC249ADCD08	15.18	06:00:00	3d 02:04:03	backfill	256	script
513593	UltrafastMat	1,341.82	1d 00:00:00	2d 02:27:31	default	4096	interactive
513598	DirectFusion	119.55	09:00:00	2d 00:23:41	default	520	script
513613	spentFuel	51.00	01:00:00	1d 23:48:13	default	256	script
513654	TurbShockWalls	106.47	1d 00:00:00	1d 16:48:06	default	832	script
513656	TurbShockWalls	105.66	1d 00:00:00	1d 16:36:07	default	832	script
513722	HHPMT_5	51.07	06:00:00	1d 13:40:04	default	256	script
513736	HierChemSep	64.65	09:00:00	1d 13:13:09	default	512	script
513758	PSFMat_2	146.57	06:00:00	1d 12:13:35	default	512	script
513759	PSFMat_2	146.44	06:00:00	1d 12:12:36	default	512	script
513760	PSFMat_2	65.85	12:00:00	1d 12:09:34	default	640	script
513768	IonTransES	124.08	06:00:00	1d 11:42:09	default	409	script
513769	IonTransES	79.53	09:00:00	1d 11:40:36	default	540	script
513771	IonTransES	87.89	09:00:00	1d 11:20:44	default	718	script
513823	HierChemSep	122.00	03:00:00	1d 08:48:33	default	512	script

Workload managers/Schedulers

- Portable Batch System (PBS)
- LoadLeveler
- Application Level Placement Scheduler (ALPS)
- Load Sharing Facility (LSF)
- Moab/Torque
- Simple Linux Utility for Resource Management (SLURM)

Batch Queueing Systems – Features

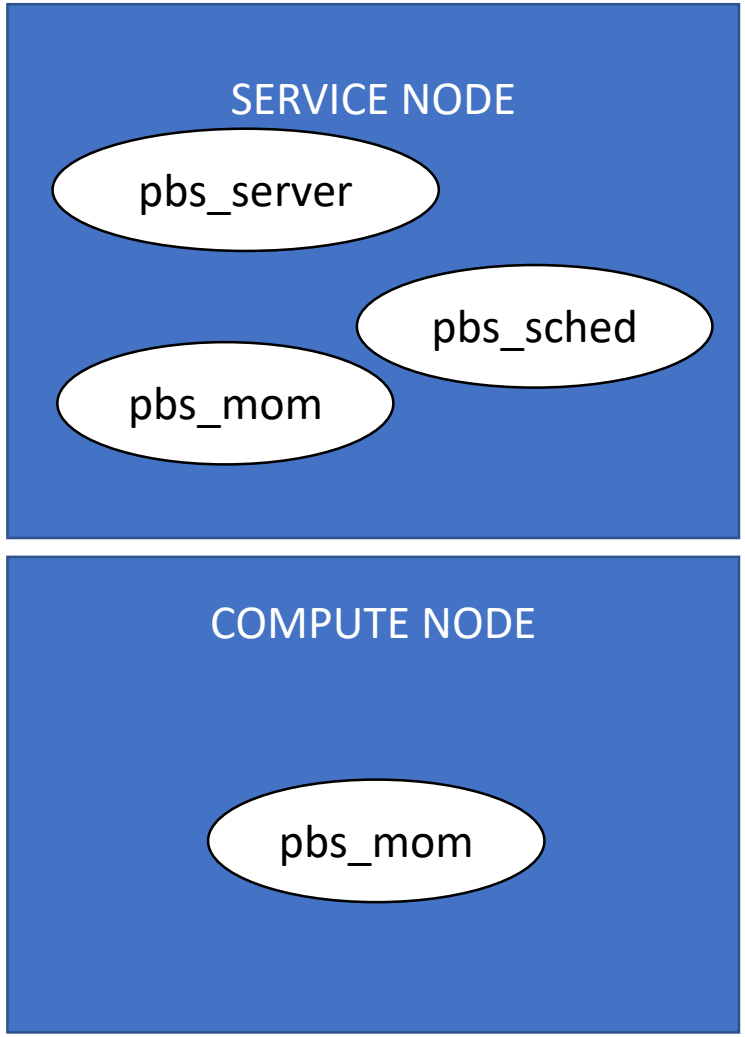
- Schedules jobs based on queues
- Has full knowledge of queued, running jobs
- Has full knowledge of the resource usage
- Typically FCFS with backfilling
- Often combination of best fit, fair share, priority-based
- Designed to be generic, can be customized
- Suited to meet demands of the scheduling goals of the centre

Portable Batch Scheduler (PBS)

- Genesis of PBS in NASA (from Network Queuing System)
- Client commands for submission, modification, and monitoring jobs
- Daemons running on service nodes, compute nodes, and servers

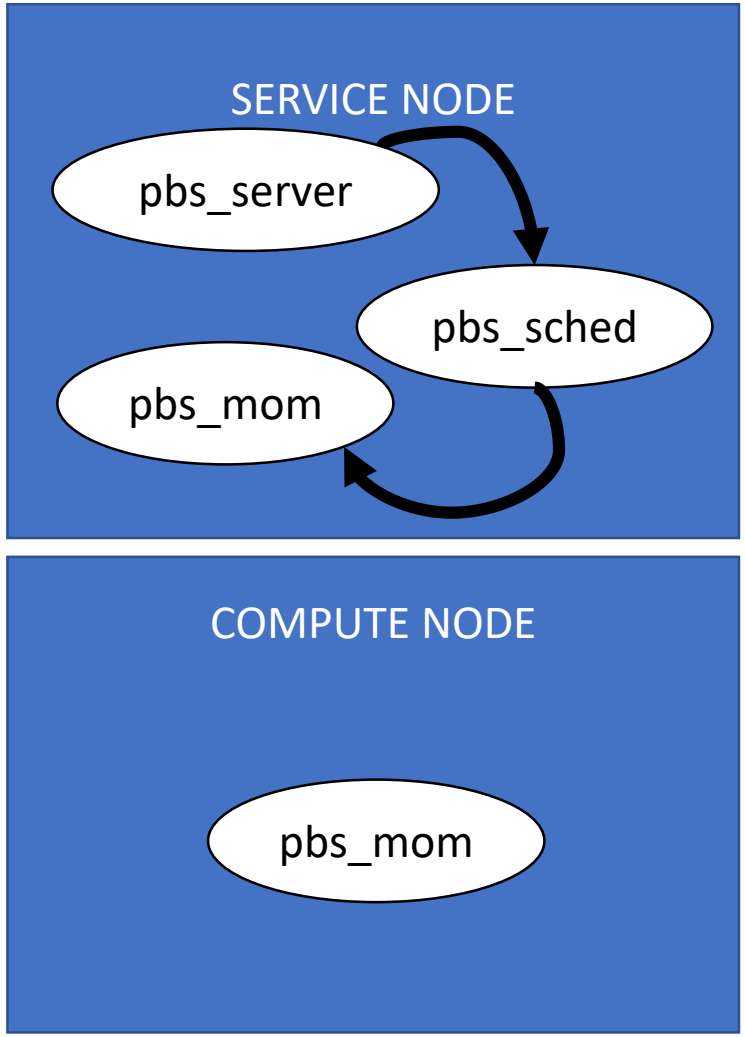
PBS daemons

- Server (pbs_server)
  - Handles PBS commands
  - Creates batch jobs
  - Sends jobs for execution
- Scheduler (pbs_sched)
  - Schedules jobs according to system policy
- MOM (pbs_mom)
  - Manage job execution on hosts
  - Resource usage monitor
  - Record diagnostic messages
  - Notify server about job completion
  - Clean up after job completion

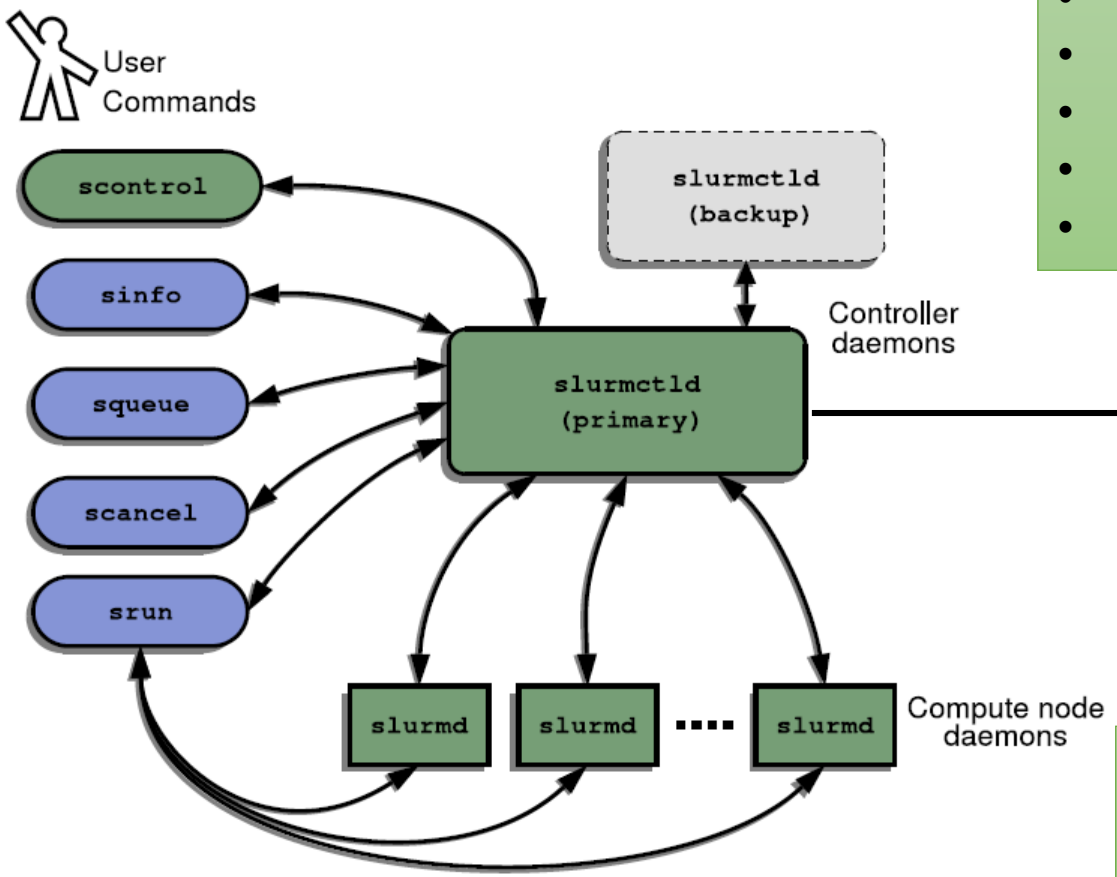


PBS daemons

- Server contacts scheduler
  - Job is queued
- Scheduler contacts the resource monitor (MOM)
  - Queries resource usage



SLURM



Slurm architecture [Jette et al.]

- Monitors states of nodes
- Accepts job requests
- Maintains queue of requests
- Schedules jobs
- Initiates job execution and cleanup
- Polls slurmd periodically
- Maintains complete state information

- Responds to controller requests
- Maintains job state
- Initiate, manage, cleanup processes
- I/O handling

An Example Scheduling Policy

1. Prime time, 6 AM to 6 PM

   a. When less than 113 nodes in use:
   1-32 node jobs limited to < 4 hours.
   >32 node jobs limited to < 10 minutes.

   b. When more than 112 nodes are already in use:
   jobs limited to < 10 minutes. This maintains high availability on the last 32 nodes.

2. Interactive Extension Period, 4 AM to 6 AM and 6 PM to 10 PM

   a. When less than 113 nodes in use:
   1 - 128 node jobs limited to < 6 hr.
   > 128 node limited to < 10 min.

   b. Jobs using last 16 nodes are limited to less than 10 minutes

   c. Jobs are not started if they might not complete before the end of the shift.

3. Night time, 10 PM to 4 AM Monday through Friday and all day Saturday and Sunday.

   a. 1-144 node jobs limited to < 6 hours.

   b. Jobs are not started if they might not complete before the end of the shift.

Henderson, “Job Scheduling Under the Portable Batch System”, JSSPP 1995.

David Lifka, The ANL/IBM SP Scheduling System, JSSPP 1995

Desirable Features of Scheduler

- Fair
- Simple
- Reduce average queue wait times
- Ensure high system utilization
- Provide optimum performance for all kinds of jobs
- Support different job classes (interactive vs. batch)
- Provide priority for special jobs

User Jobs

- Different types of jobs: task farms vs. message passing
- Interactive vs. batch jobs
  - Debug in interactive mode
- Exclusive vs. shared access
- Charged based on total resource usage
  - Job is killed when requested wall-clock time is over
  - Need to plan resource usage apriori

Scheduler Queues

- Jobs are submitted to a queue
- Different queuing policies (decided by the administrator)
- Multiple queues in some systems
  - Based on the usage
  - Queue waiting time different
  - Static vs. dynamic partitioning

FCFS with Backfilling

- FCFS scheduling
  - Poor system utilization
- Backfilling – to overcome inefficiency of FCFS
- Scan the queue of jobs for a job that does not cause the first queued job to wait for any longer than they otherwise would
- Improve system utilization
- Lower queue waiting times

Backfilling – Example

User Name	Number of Nodes	Number of Minutes	Job Status
User A	32	120	Startable
User B	64	60	Waiting
User C	24	180	Waiting
User D	32	120	Waiting
User E	16	120	Waiting
User F	10	480	Waiting
User G	4	30	Waiting
User H	32	120	Waiting

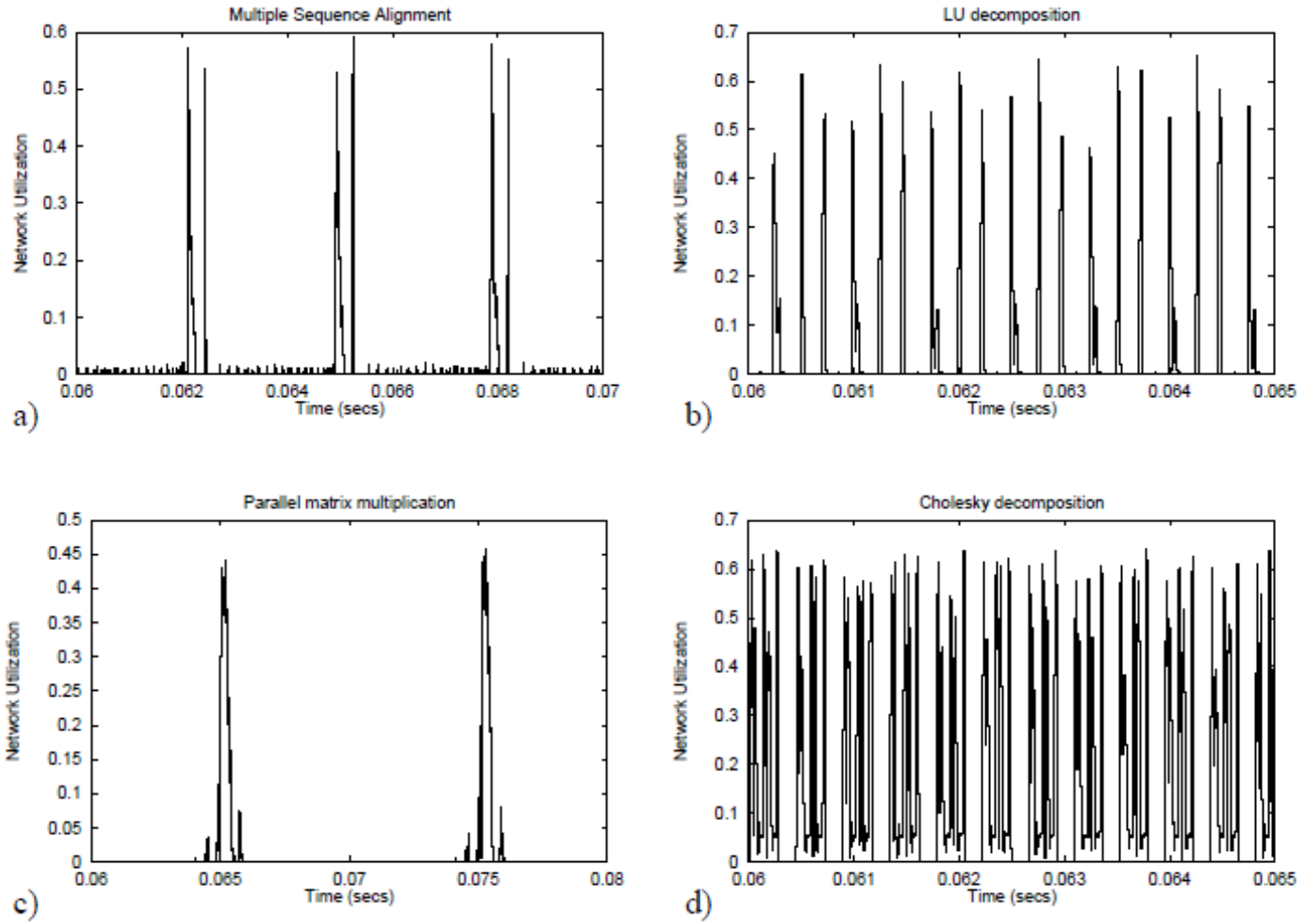
User Name	Number of Nodes	Number of Minutes	Job Status
User A	32	120	Running
User B	64	60	Startable
User C	24	180	Waiting
User D	32	120	Waiting
User E	16	120	Waiting
User F	10	480	Waiting
User G	4	30	Waiting
User H	32	120	Waiting

User Name	Number of Nodes	Number of Minutes	Job Status
User A	32	120	Running
User B	64	60	Running
User C	24	180	Running
User D	32	120	Blocked
User E	16	120	Ineligible
User F	8	480	Startable
User G	4	30	Waiting
User H	32	120	Waiting

User Name	Number of Nodes	Number of Minutes	Job Status
User A	32	120	Running
User B	64	60	Running
User C	24	180	Startable
User D	32	120	Waiting
User E	16	120	Waiting
User F	10	480	Waiting
User G	4	30	Waiting
User H	32	120	Waiting

User Name	Number of Nodes	Number of Minutes	Job Status
User A	32	120	Running
User B	64	60	Running
User C	24	180	Running
User D	32	120	Blocked
User E	16	120	Ineligible
User F	10	480	Ineligible
User G	4	30	Startable
User H	32	120	Waiting

Network Utilization in Different Applications



Petrini and Feng, Time-Sharing Parallel Jobs in the Presence of Multiple Resource Requirements, JSSPP 2000