The Role of MPPs at NASA and the Aerospace Industry

Manny Salas

May 22, 1997

Talk Outline

- NASA Aeronautics Computational Resources
- NASA Utilization & Availability Metrics
- NASA Aeronautics Parallel Software
- Aerospace Industry

NASA Aeronautics Computational Resources

http://www.nas.nasa.gov/aboutNAS/resources.html

- **Aeronautics Consolidated Supercomputing Facility (ACSF)** is a shared resource for supercomputing for Ames, Dryden, Lewis and Langley. It is located at NASA Ames.
- **Numerical Aerospace Simulation Facility (NAS)**

  The NAS Facility provides pathfinding research in large-scale computing solutions. NAS serves customers in government, industry, and academia with supercomputing software tools, high-performance hardware platforms, and 24-hour consulting support. NAS is also located at NASA Ames.

An Overview of ACSF Resources

- A CRAY C90 (Eagle) consisting of an 8-processor system with 256 MW (64 bits/word) of central memory and 512 MW of SSD memory. The CRAY C90 has a peak performance of 1 GFLOPS per CPU.
- A cluster of four SGI/Cray J90 systems (Newton). The configuration consists of one 12 processor J90SE (512 MW), one 12 processor J90 (512 MW), one 8 processor J90 (128 MW), and one 4 processor J90SE (128 MW) systems.

An Overview of NAS Resources and Services

- The IBM SP2 (Babbage) is a 160 node system based on RS6000/590 workstations. Each node has 16 MW of main memory and at least 2 GB of disk space. The nodes are connected with a high-speed, multi-stage switch. Peak system performance is over 42.5 GFLOPS.
- The CRAY C90 (von Neumann) is a 16-processor system with 1 GW (64 bits/word) of central memory and 1 GW of SSD memory. The CRAY C90 has a sustained aggregate speed of 3 GFLOPS for a job mix of computational physics codes. Peak performance is 1 GFLOPs per CPU.

An Overview of NAS Resources and Services

- SGI Origin 2000 (Turing) consisting of 2-32 node systems configured as a 64 CPUsystem with 2 GW. Each CPU delivers about 400 MFLOPS.
- The NAS HPCC workstation cluster, an SGI PowerChallenge Array (Davinci), consists of one front end system and eight compute nodes. The front end system is the host that users log into. The front end is an SGI PowerChallenge L with four 75 MHZ R8000 CPUs and 48MW memory, and serves as the system console, compile server, file server, user home server, PBS server, etc. There are eight compute nodes (four two-CPU nodes & four eight-CPU shared-memory nodes).

An Overview of NAS Resources and Services

- Two 2-processor Convex 3820 mass storage system
- Silicon Graphics 8-processor 4D/380S support computers
- Silicon Graphics, Sun, IBM and HP workstations
- All the machines are currently connected via Ethernet, FDDI, and HiPPI. ATM network adapters from both SGI and Fore Technology are currently being tested

An Overview of NAS Resources and Services

File Storage:

- 350 GB on-line CRAY C90 disk
- 1.6 TB on-line mass storage disk
- 5 Storage Technology 4400 cartridge tape robots (50 TB of storage)

HPCCP Resources at Goddard

A Cray T3E with 256 processors. Each processor has 128 MB of memory, 32 GB total. Peak performance 153 GFLOPS. An additional 128 processors and 480 GB disk to be installed in June with peak performance of 268 GFLOPS.

Summary of Resources

Program	Type	Name	Processors	Memory GW	Comp. rate Gflops
ACSF	CrayC90	Eagle	8	.256	8
ACSF	Cray J90 cluster	Newton	36	1.28	16
NAS HPCC	IBM SP2	Babbage	160	2.5	42.5
NAS	CrayC90	v.Neumann	16	1	16
NAS	SGI Origin	Turing	64	2	25
NAS HPCC	SGI PC array	Davinci	44	1.7	3
HPCC Goddard	CrayT3E		384	6	268

Utilization & Availability Metrics

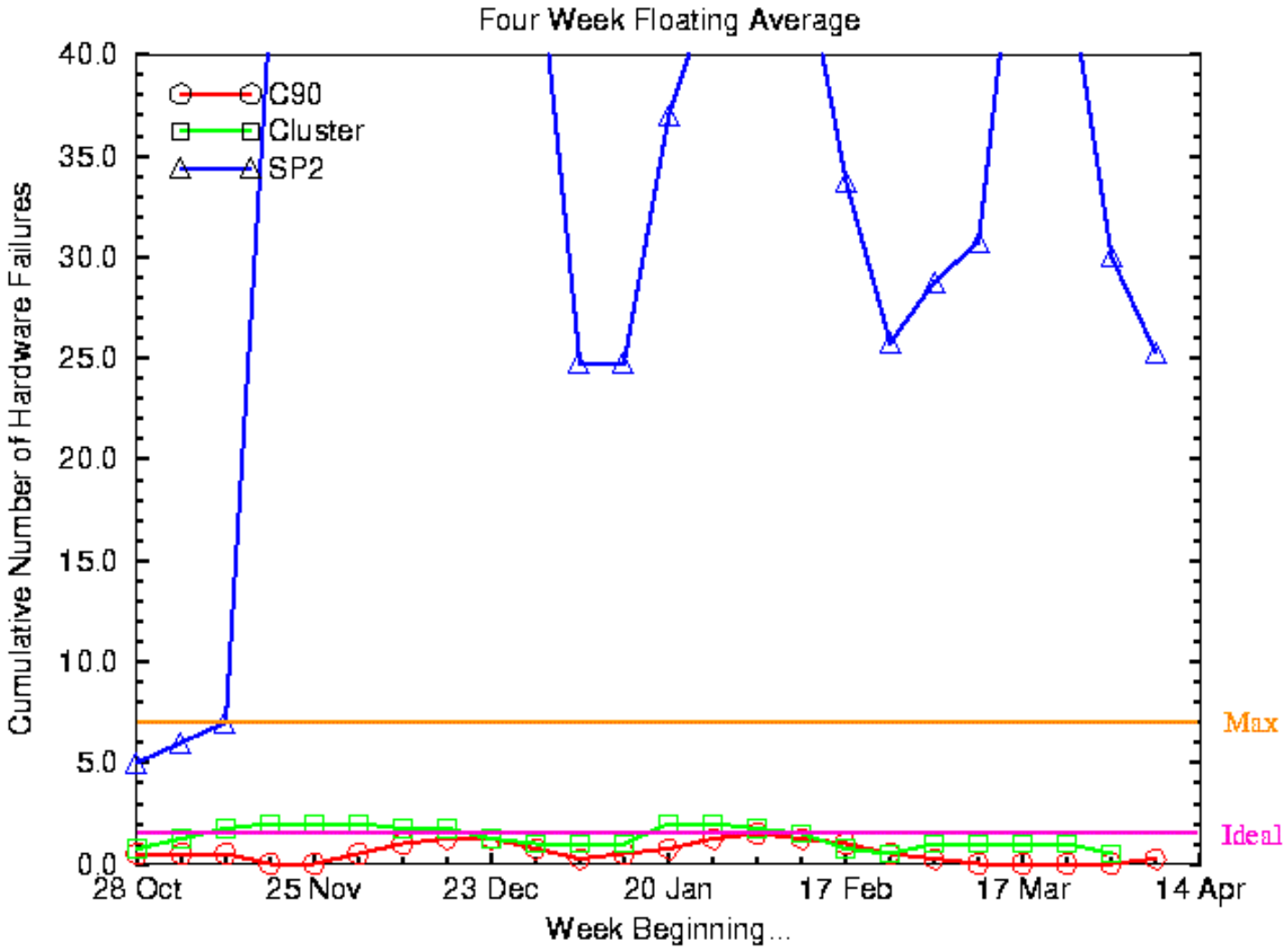
http://wk122.nas.nasa.gov/HPCC/Metrics/metrics.html

http://www-parallel.nas.nasa.gov/Parallel/Metrics/

Number of Unscheduled Interrupts

Counts the number of unscheduled service interruptions. An "interrupt" in this context could be a software failure (crash), hardware failure, forced reboot to clear a problem, facility problem, or any other unscheduled event that prevented a significant portion of the machine from being useful. The maximum acceptable value for each system is one per day, and the goal is less than one per week.

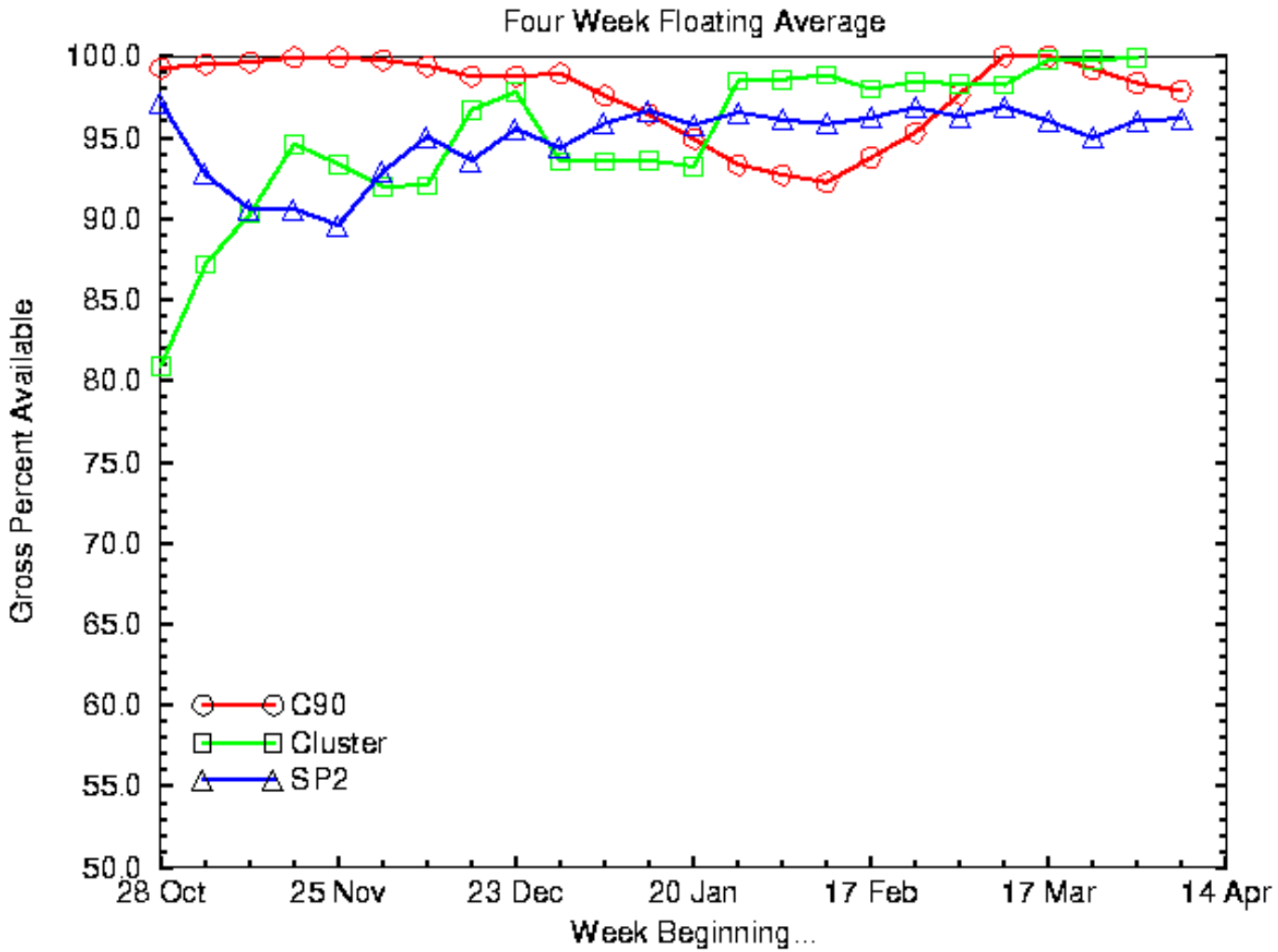
Number of Unscheduled Interrupts



Gross Availability

Shows the fraction of time that the machine was available, regardless of whether it was scheduled to be up or down. Dedicated time reduces this number but has no effect on "net availability". Acceptable values and goals are not defined.

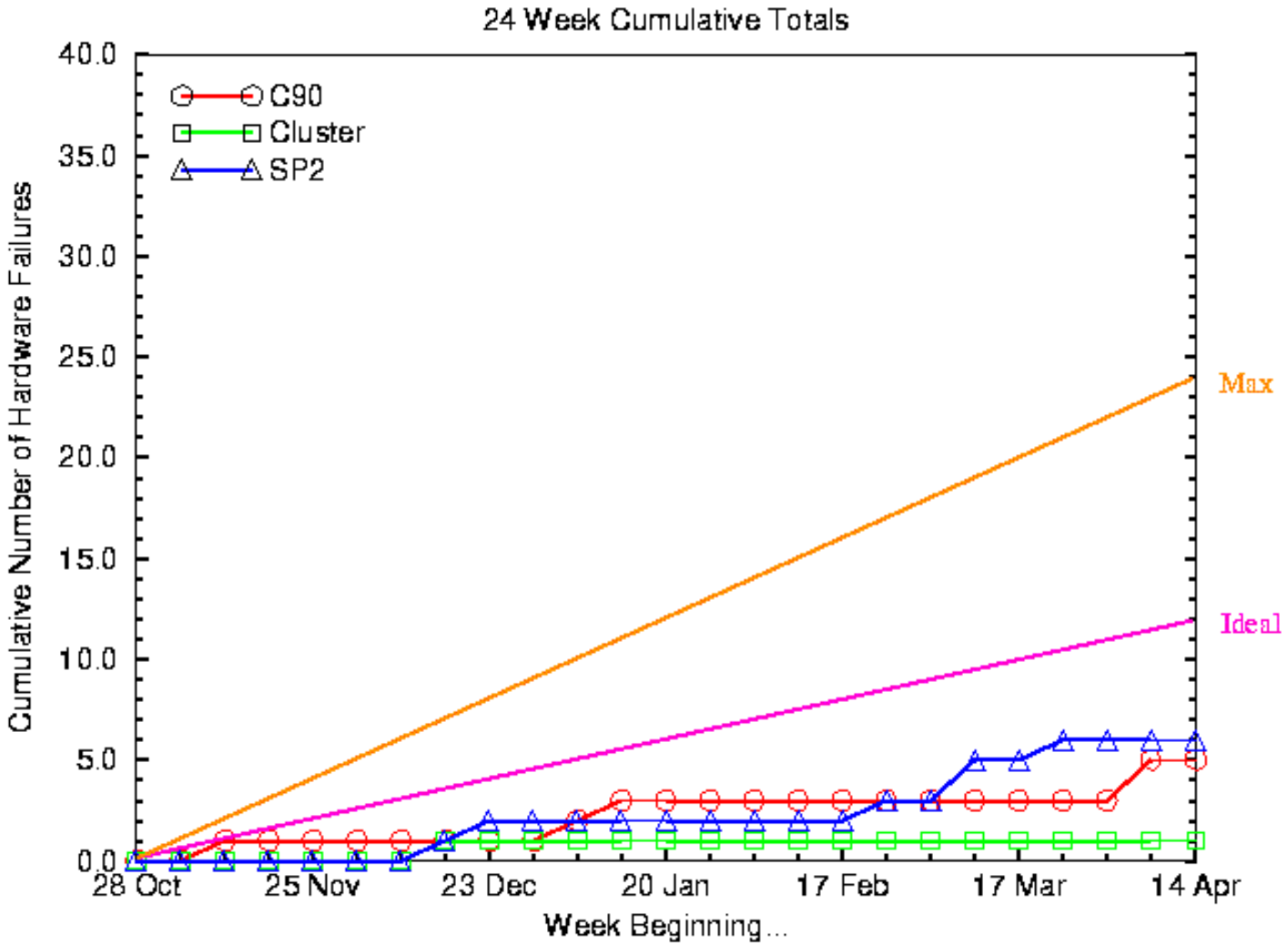
Gross Availability



Hardware Failures

Count of hardware faults that required physical repair or replacement. Hardware failures do not necessarily result in an interrupt, because some machines (C90) perform error correction and can continue to run with damaged hardware. The replacement would then take place in scheduled maintenance time, counting as a hardware failure but not as an interrupt. The maximum acceptable value is one hardware failure per week, and the goal is less than one per month.

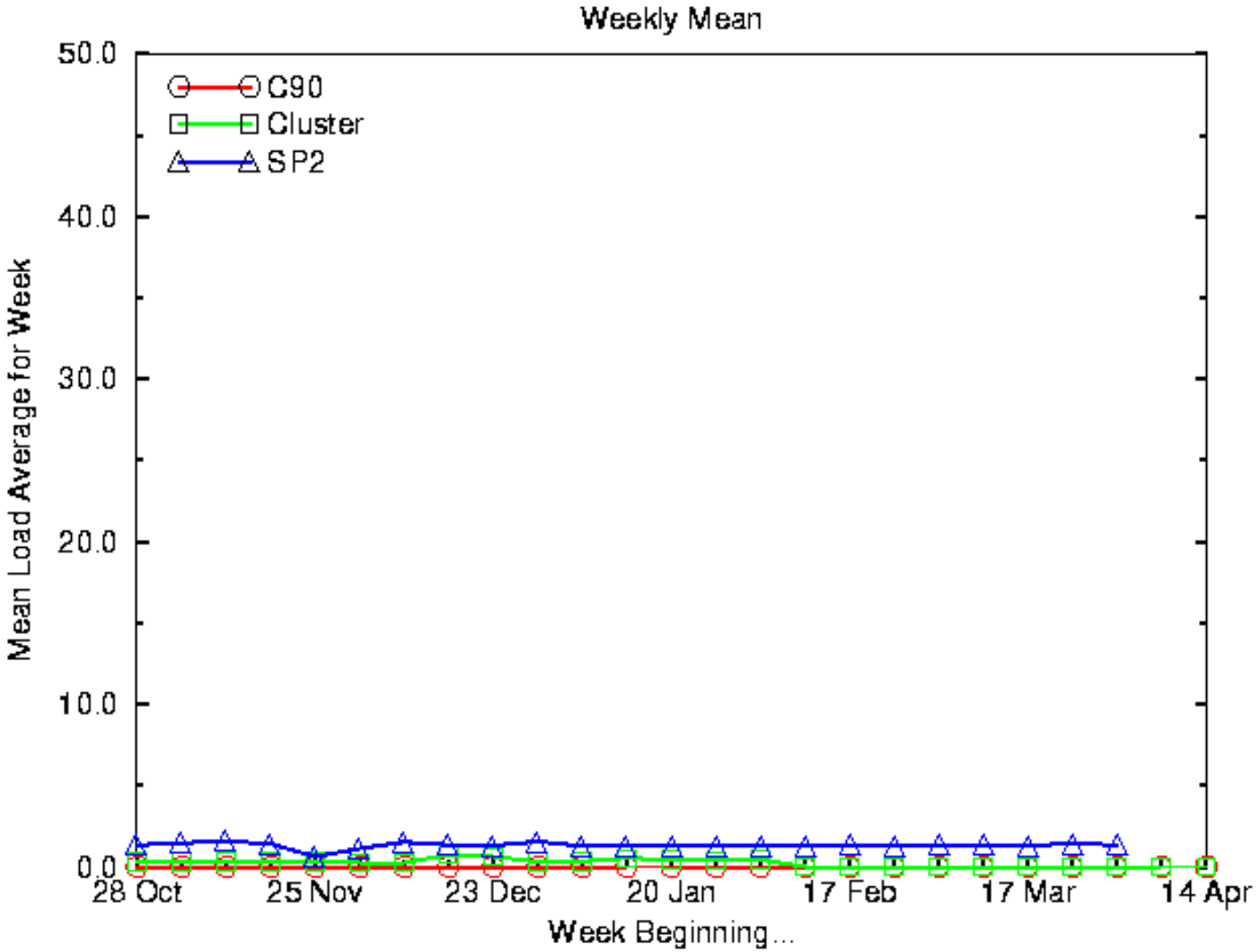
Hardware Failures



Load Average

Mean "load average" for a week. Sampled every fifteen minutes throughout the week and averaged. "Load average" is the average number of processes that are ready to run, reflecting the busy-ness of the machine. There are no goals yet set, and for most parallel systems, time-sharing is still a bad idea, hence the ideal load average should be close to 1 (or 1 times the number of compute nodes, if the load average is not normalized).

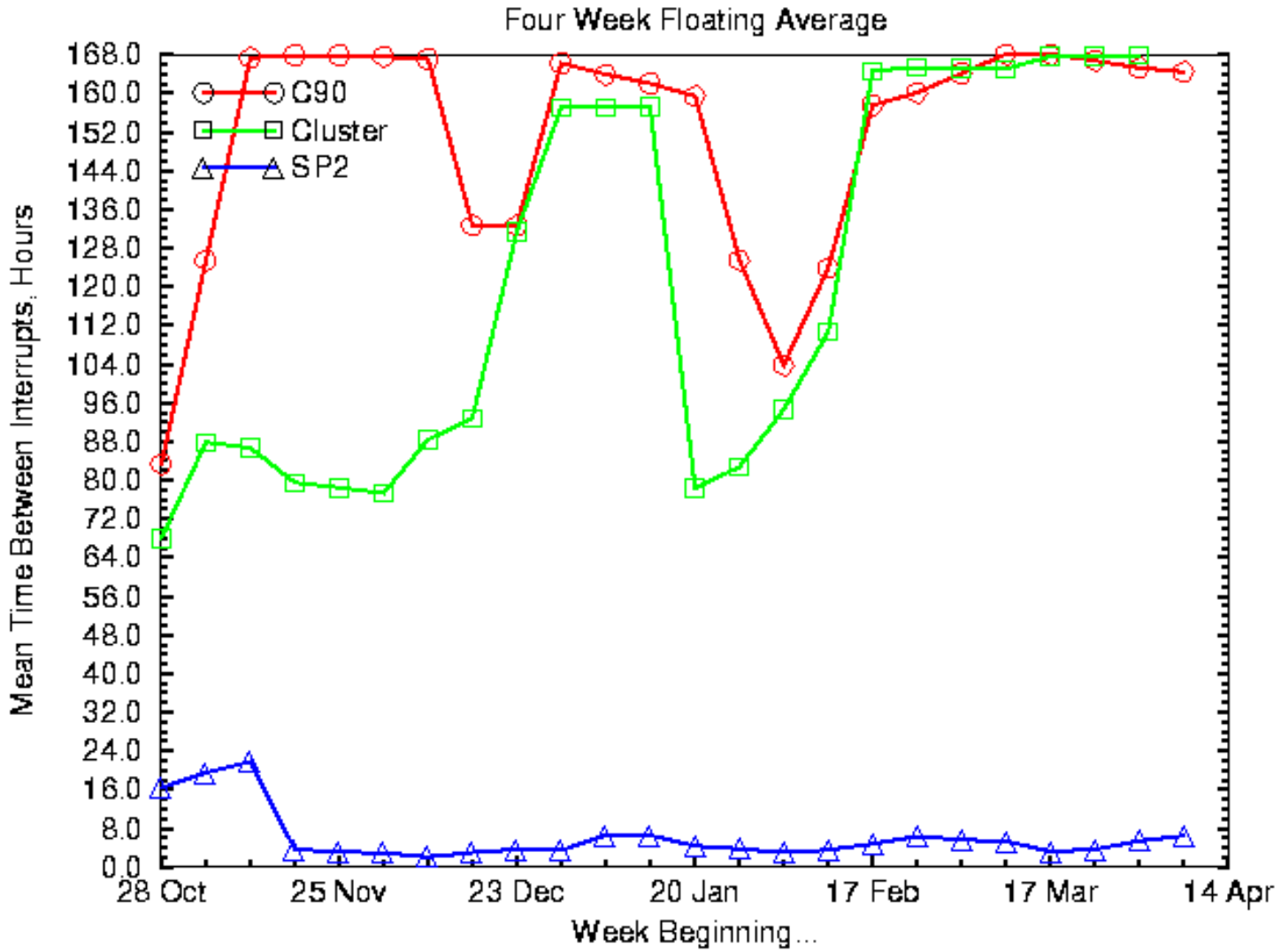
Load Average



Mean Time Between Interrupts

Mean time between unscheduled service interruptions. Calculated by dividing the uptime by the number of interruptions plus one. Acceptable values and goals for this metric are not separately defined, but are dependent on the values chosen for the "Number of Unscheduled Interrupts" and "Mean Time To Repair" metrics.

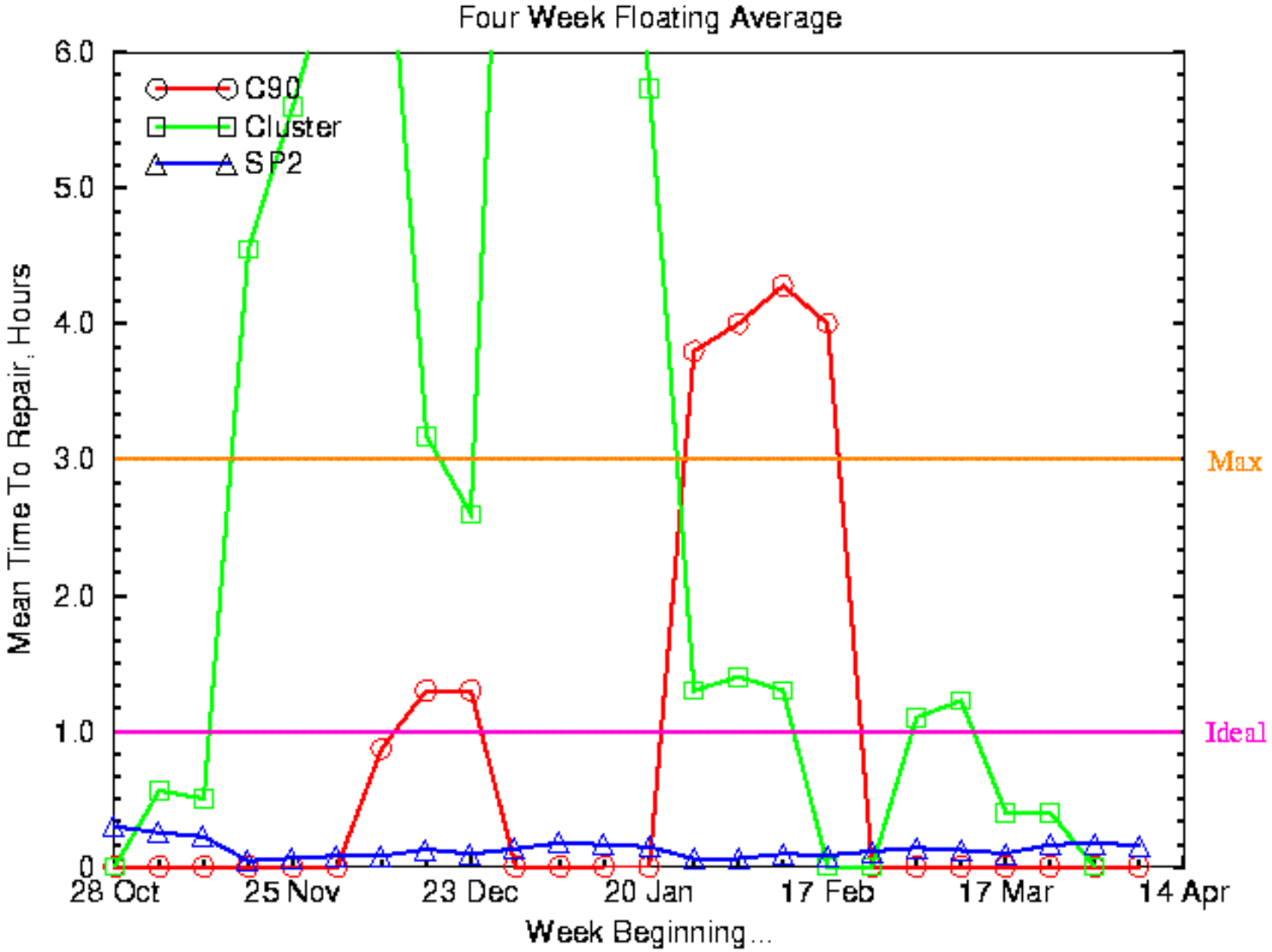
Mean Time Between Interrupts



Mean Time To Repair

Average time required to bring the machine back up after an unscheduled interrupt. The outage begins when the machine is recognized to be down and ends when it begins processing again; it is not a reflection of the vendor FE response time. Maximum acceptable value is 4 hours, and the goal is less than one hour.

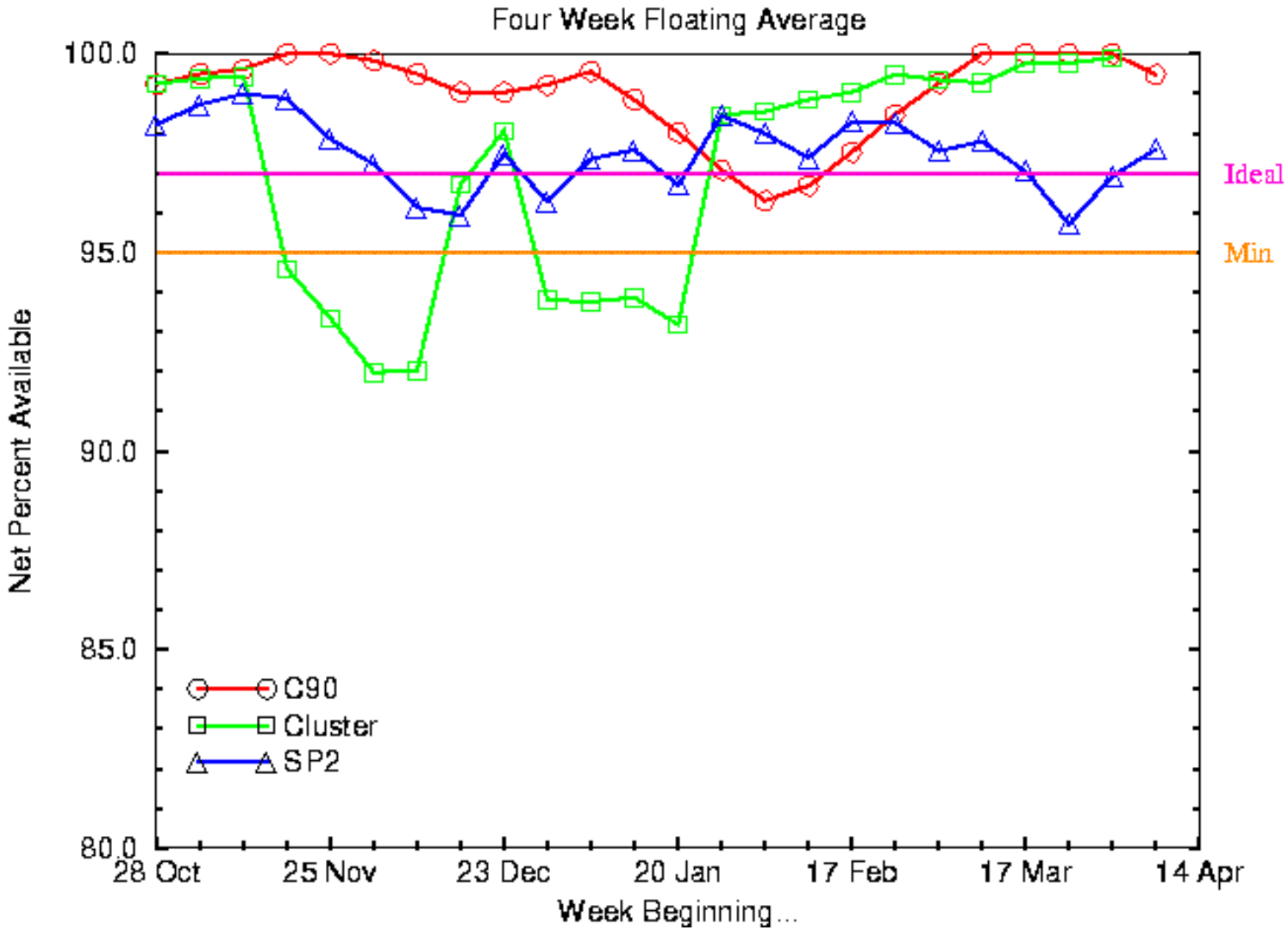
Mean Time to Repair



Net Availability

Shows the time that the machine was available as a percentage of the time that the machine was scheduled to be available. The minimum acceptable value is 95%, and the goal is 97%.

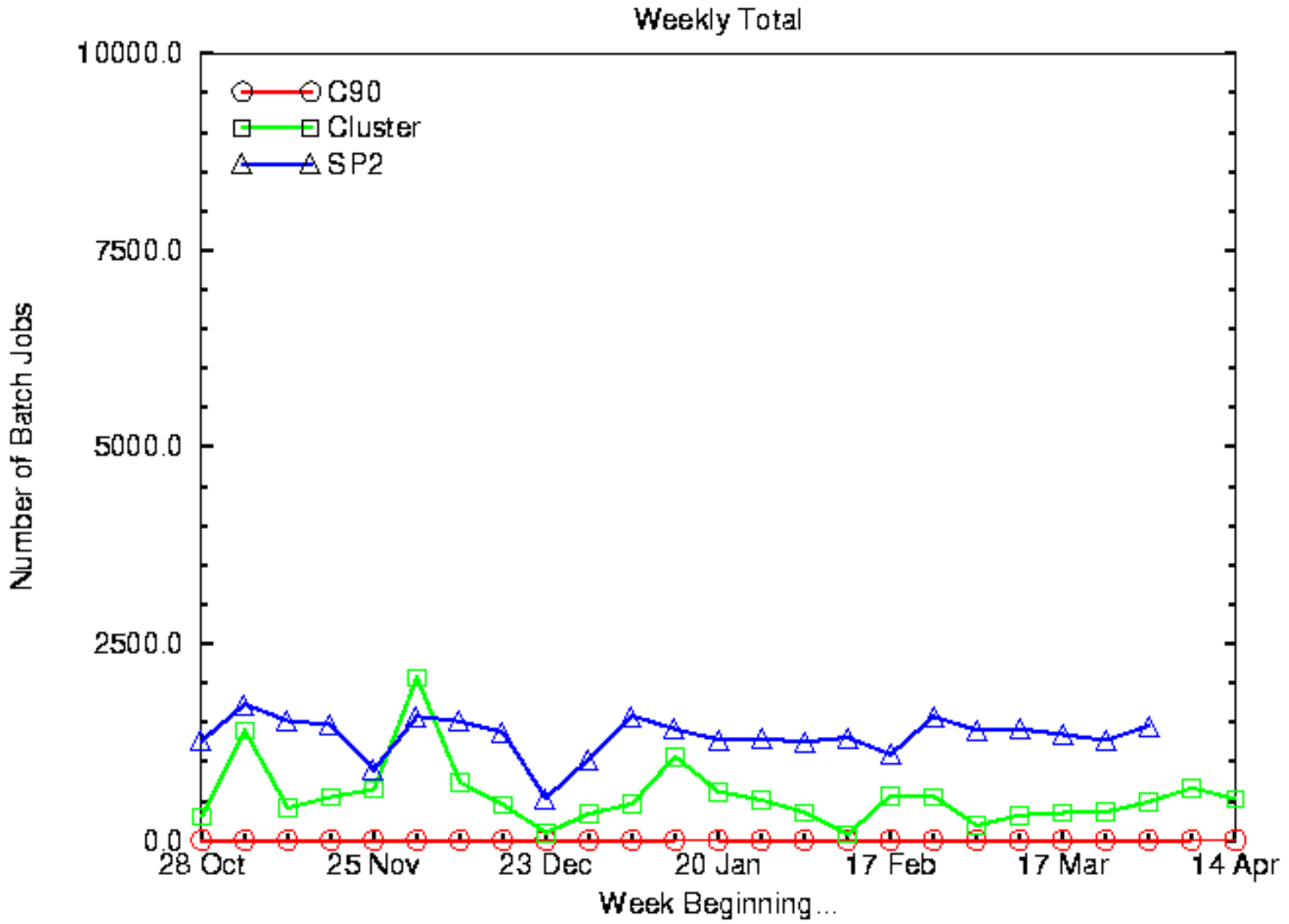
Net Availability



Number of Batch Jobs

Number of jobs run through the batch subsystem during each week. Not all jobs, even long jobs, are run though a batch scheduler on all systems. There are no defined goals for this metric.

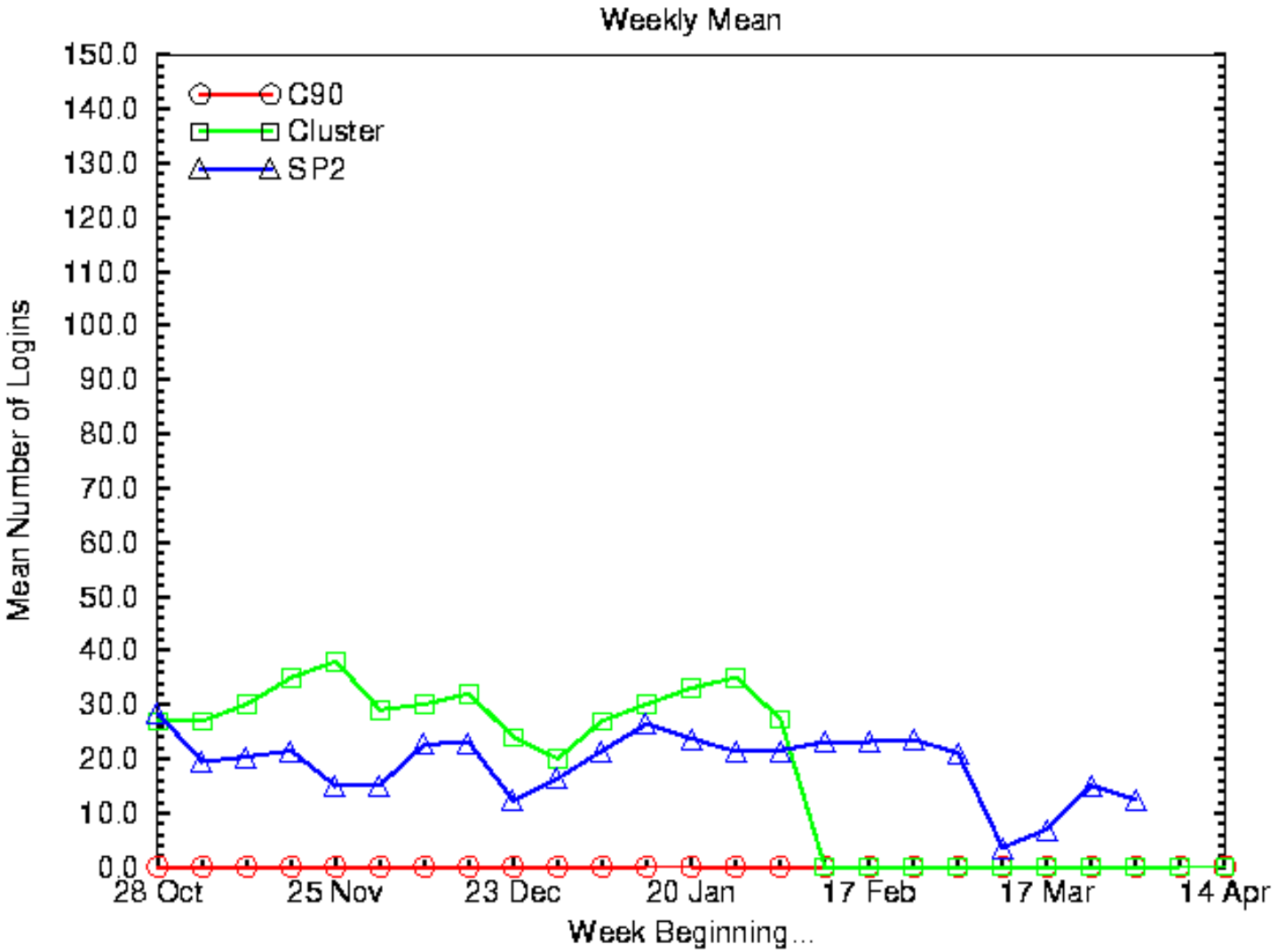
Number of Batch Jobs



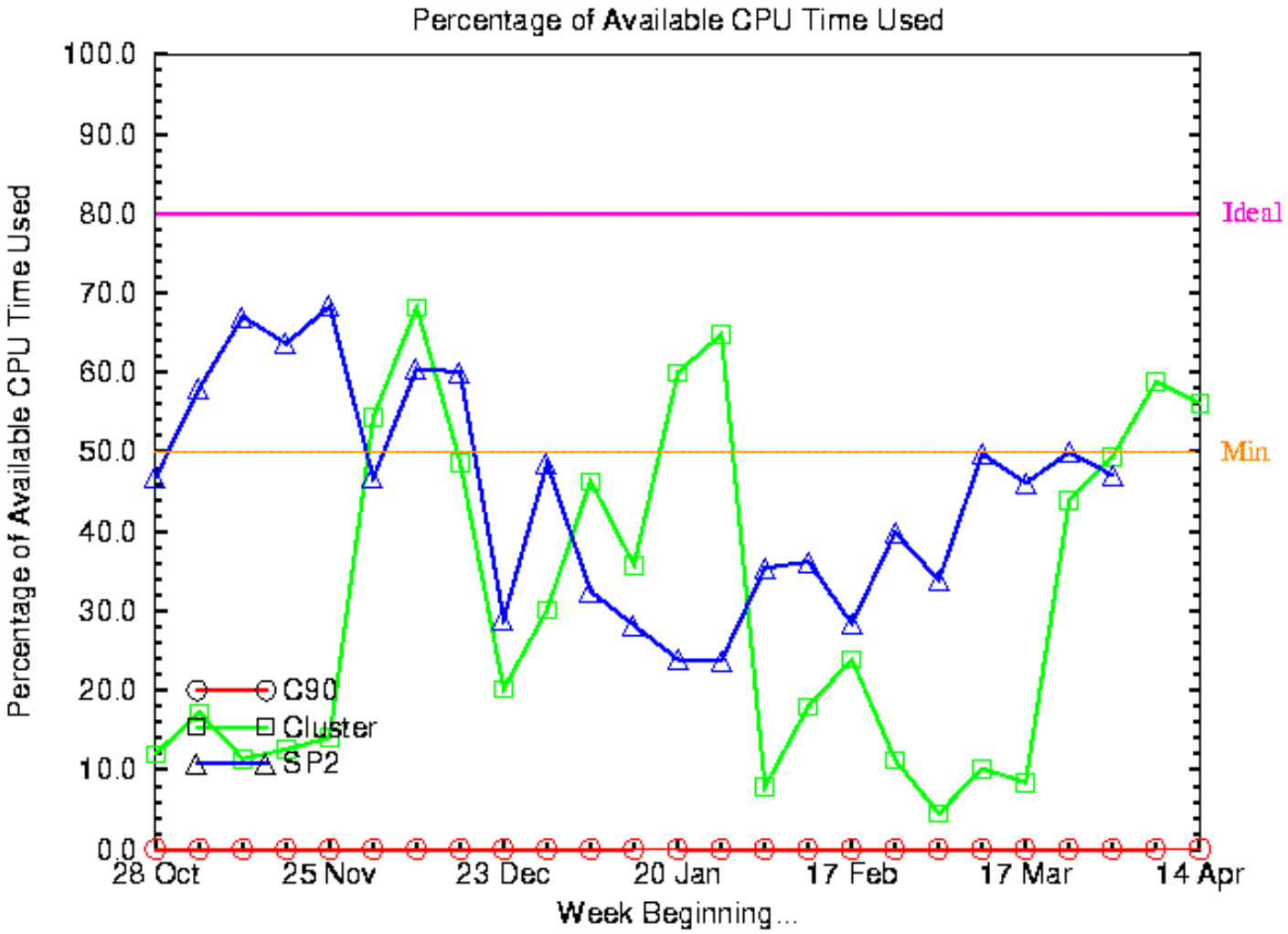
Number of Logins

Number of people logged in to the machine in question. Sampled every fifteen minutes throughout the week and averaged. This metric has no goals specified (for purely batch systems, a heavily used system may have few, or no, logins).

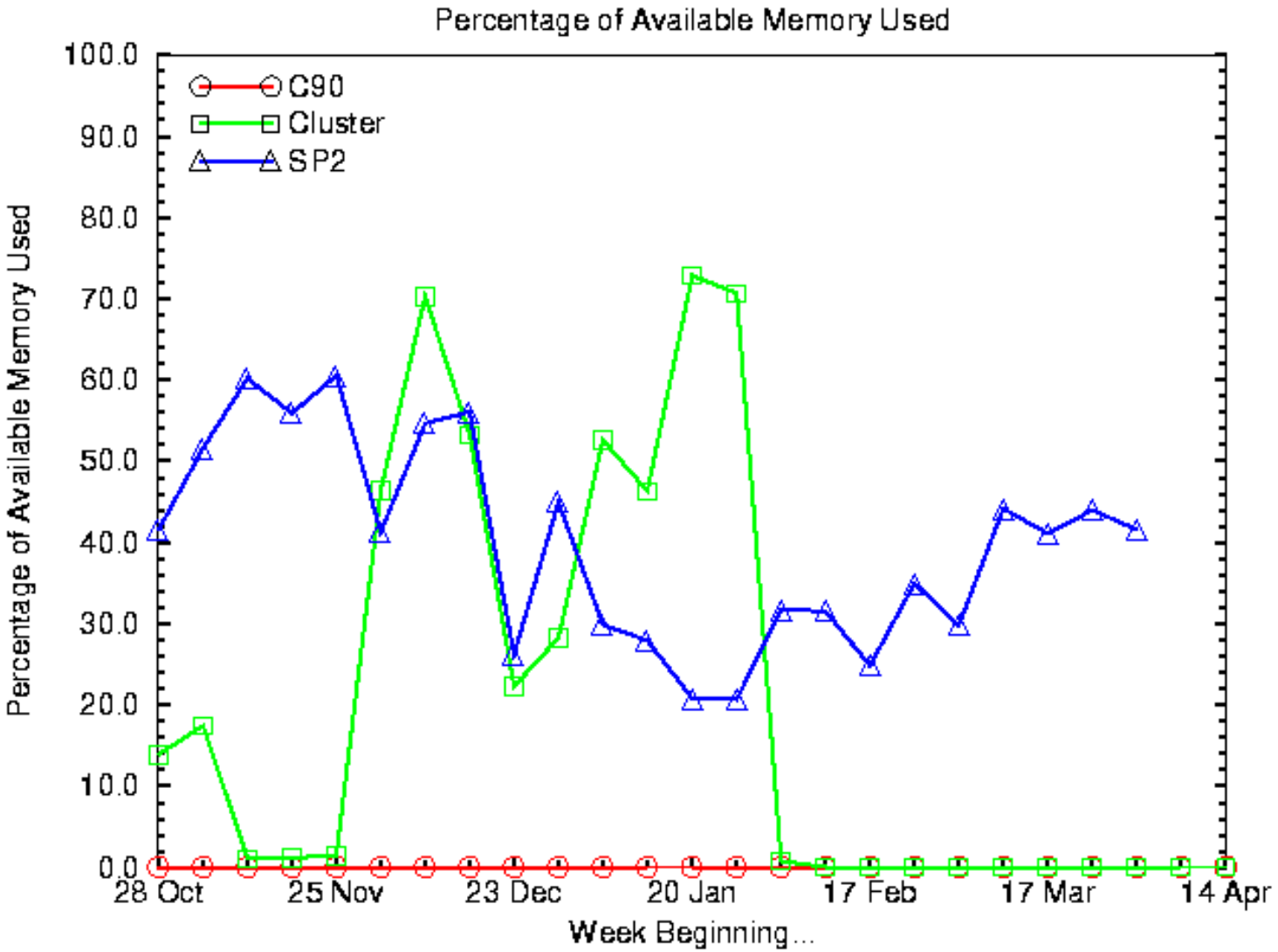
Number of Logins



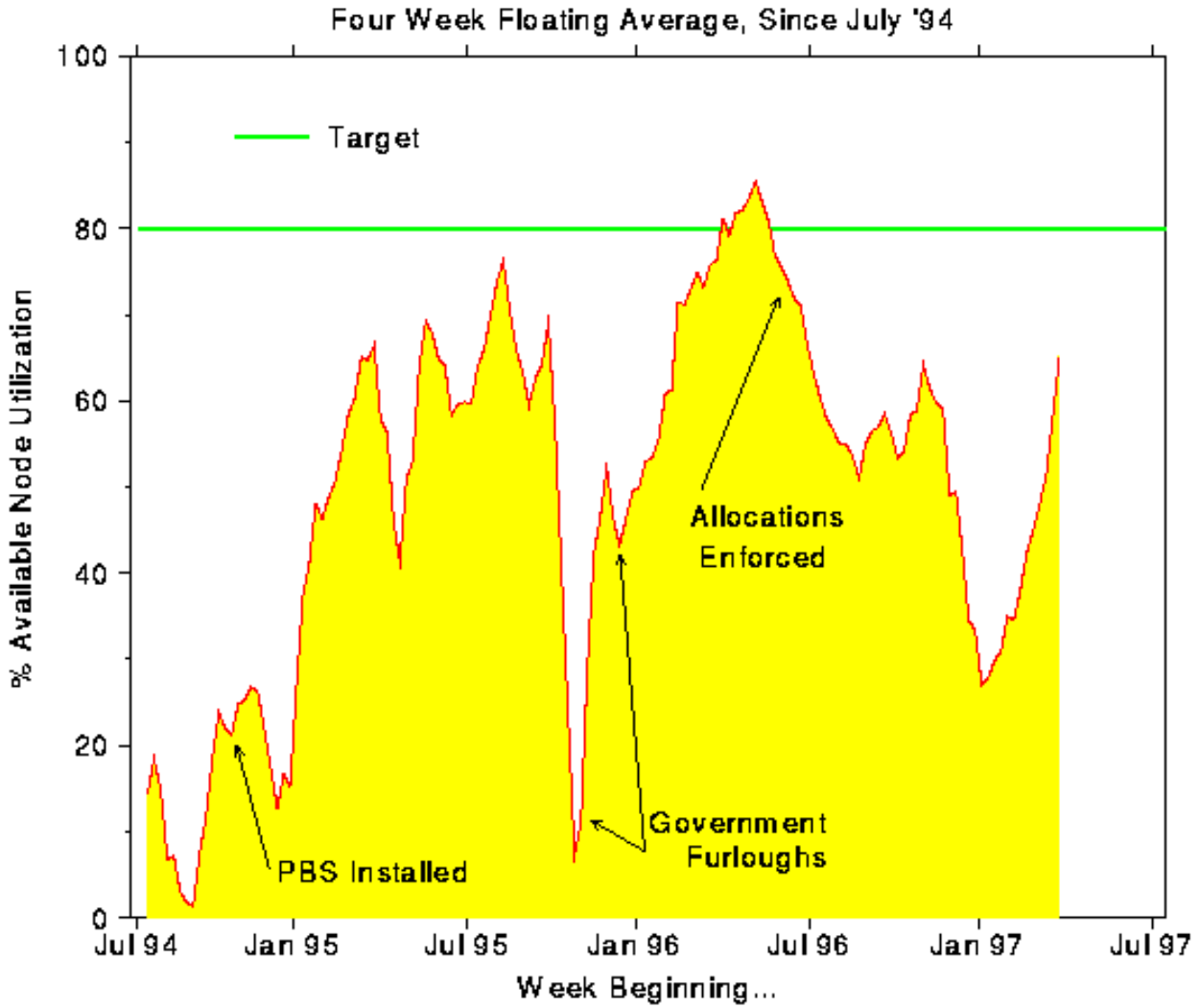
CPU Utilization



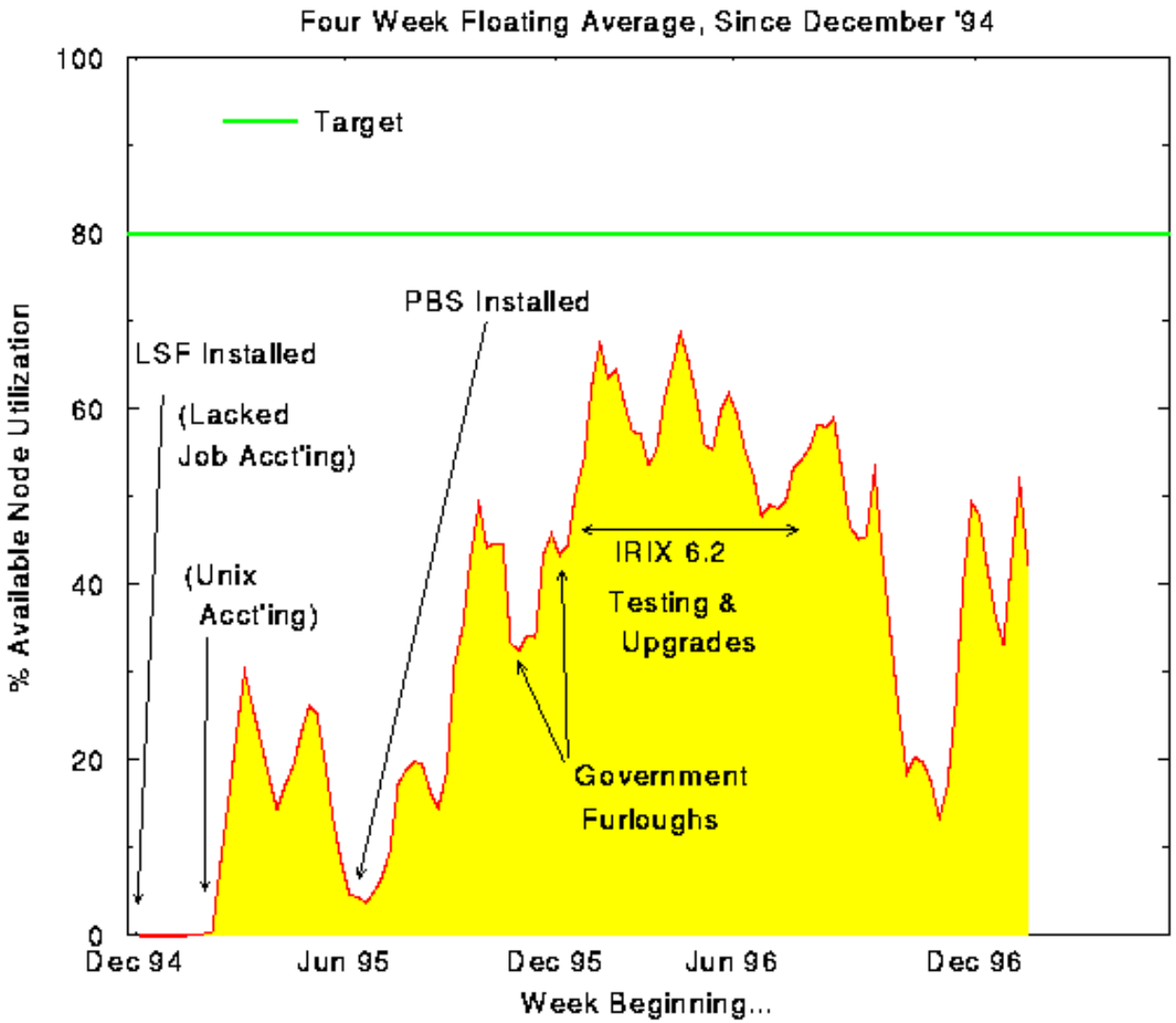
Memory Utilization



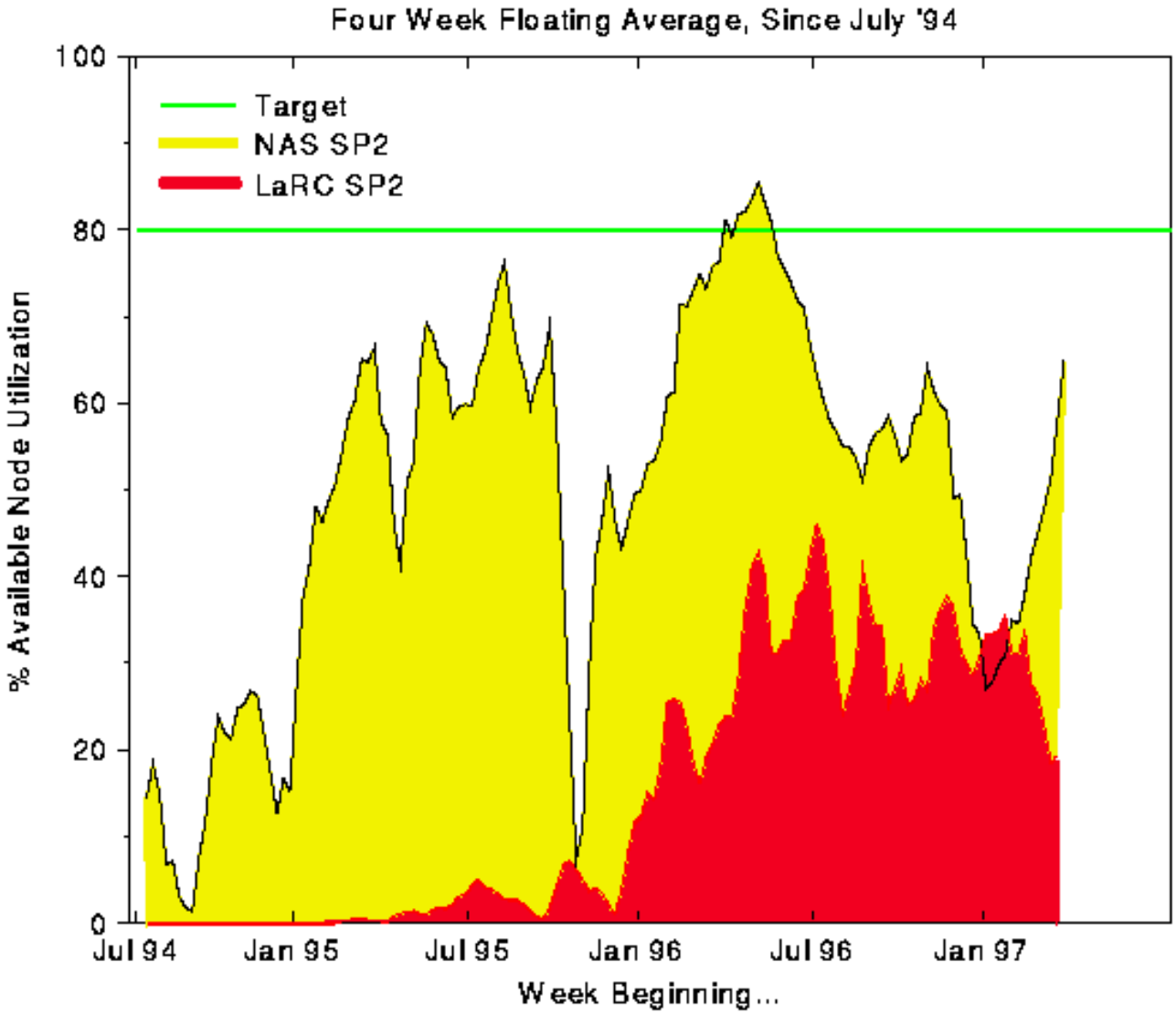
NAS HPCCP SP2 (Babbage) Utilization



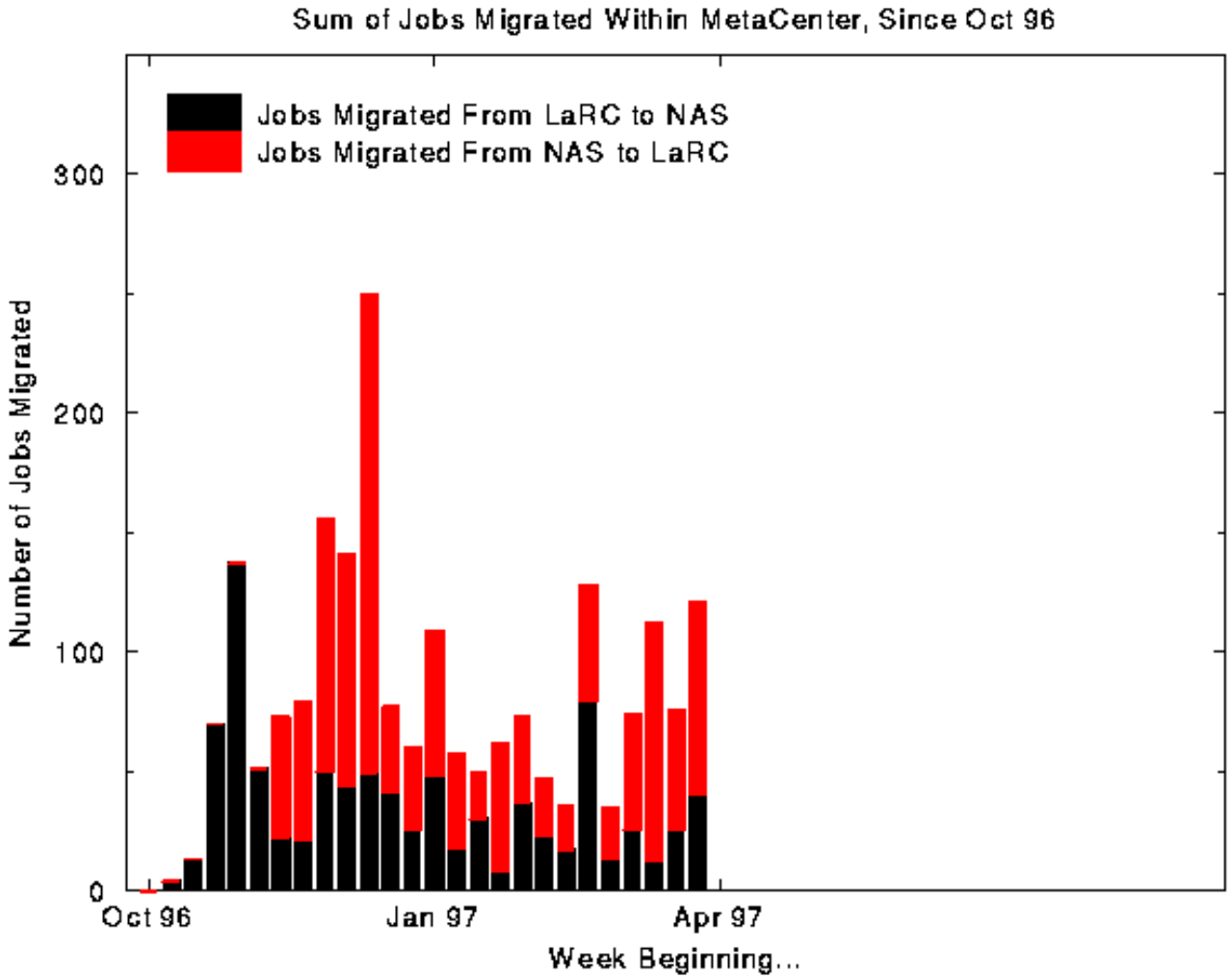
NAS HPCCP Cluster (Davinci) Utilization



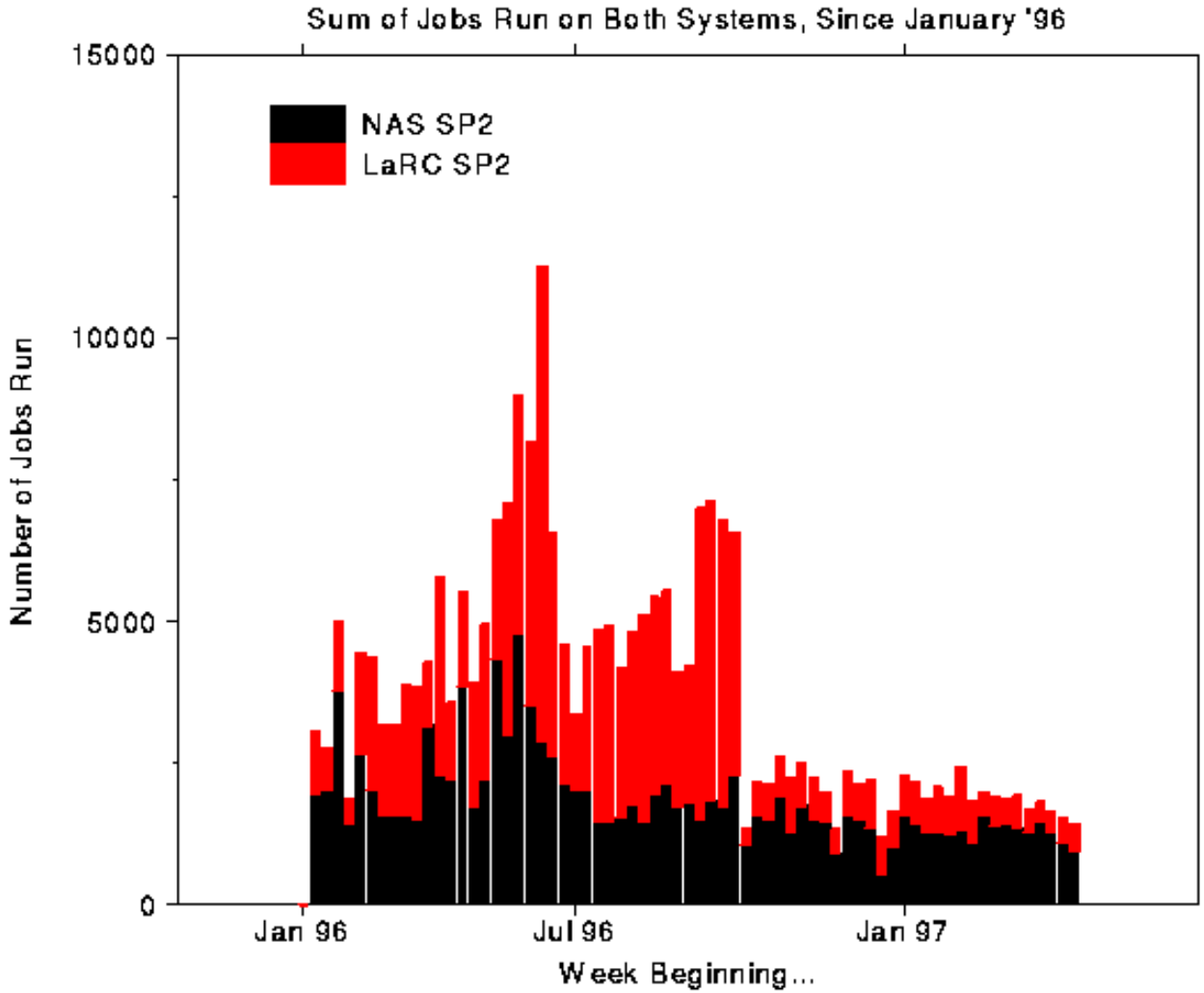
Meta Center SP2 Utilization



Meta Center SP2 Job Migration



Meta Center SP2 Batch Jobs



NASA Aeronautics Parallel Software

http://www.aero.hq.nasa.gov/hpcc/cdrom/content/reports/annrpt96/cas96/cas96.htm

NASA Codes

- **OVERFLOW**
  - Reynolds Averaged Navier-Stokes Equations
  - Overset grid topology
  - Beam-Warming Implicit Algorithm
  - Coarse grained parallelism using PVM
  - 7 SP2 processors ~ 1 C90
  - 30 workstations ~ 1 C90

NASA Codes

- **ENSAERO**
  - RANS (Overflow) + Aeroelasticity
  - Rayleigh-Ritz for elasticity
  - decomposed by discipline; 1 node for elasticity other nodes used for fluids

NASA Codes

- **CFL3D**
  - RANS (60,000 lines)
  - Implicit in time, upwind diffence, multigrid
  - multiblock & overset grids, embedded grids
  - serial & parallel executables from same code
  - 10.5 SP2 processors ~ 1 C90

NASA Codes

- **CFL3D**

No. of SP2 processors	time/(time Cray YMP)
2	1.92
4	1.00
8	.56
16	.27

NASA Codes

- **FUN3D**
  - Incompressible Euler
  - 2nd-order Roe Scheme, unstructured grid
  - Parallelized using PETSc (Keyes, Kaushik, Smith)
  - Newton-Krylov-Schwarz with block ILU(0) on subdomains

NASA Codes

- **FUN3D** tetrahedral grid with 22677 gridpoints (90708 unknowns)

SP2 Procs	iterations	exec time	speedup	efficiency
1	25	1745.4s	-	-
2	29	914.3s	1.91	.95
4	31	469.1s	3.72	.93
8	34	238.7s	7.31	.91
16	37	127.7s	13.66	.85
24	40	92.6s	18.85	.79
32	43	77.0s	22.66	.71
40	42	65.1s	26.81	.67

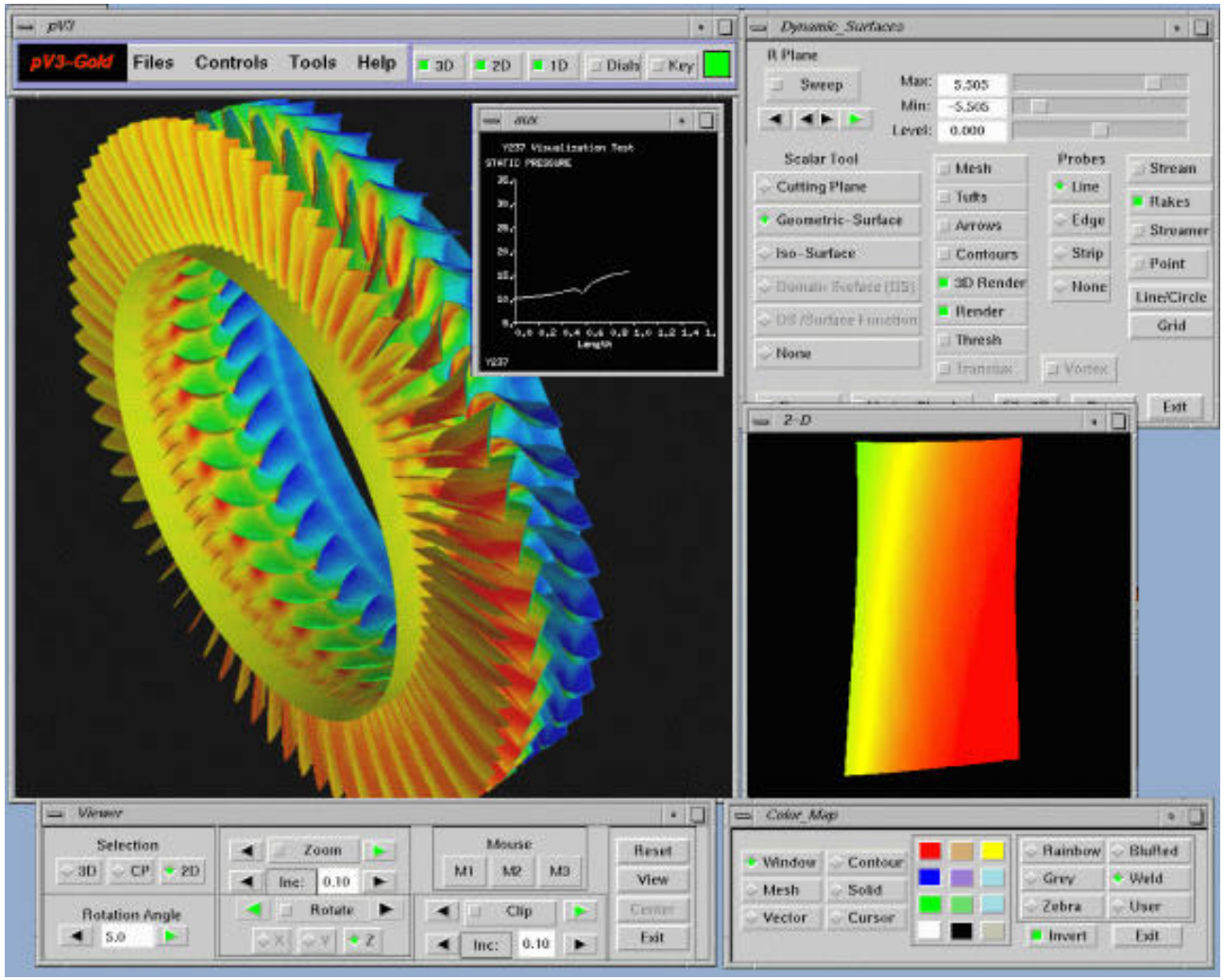
Speedup=(exec time on 1 proc)/(exec time on n procs)
Efficiency=speedup/(# of procs)

Aerospace Industry

Quote from a senior aerospace engineer at a leading aerospace company:

...we at industry are very desperate in having a fast convergent code that can solve real problems... The state-of-the-art 3D Navier-Stokes code takes in the order of 200 hours C90 time for solving one case of a high-lift application. This is *way too high*, considering we'd like to have a whole design cycle within five days...

Affordable High Performance Computing Cooperative Agreement- UTRC



Affordable High Performance Computing Cooperative Agreement

http://danville.res.utc.com/AHPCP/movie.htm

United Technologies Research Center is developing a multi-cluster environment focused on the simulation of a high-pressure compressor. The goal is to achieve overnight turnaround.

- Job migration & dynamic job sched. 12/31/96
- HP Distributed File System 3/31/97
- Multi-cluster scheduling across WANs 6/30/97
- Full system demo. 9/30/97

Aerospace Industry

- **McDonnell Douglas**
  - Extensive usage of workstation clusters, coarse grain parallelism, 100's of workstations
- **Boeing**
  - Some usage of NAS cluster (Davinci), little inhouse capability
- **Northrop Grumman**
  - Workstation cluster usage increasing, coarse grain parallelism