The Sustained System Performance Measure

Adrian Wong and Bill Kramer

November 25, 2002

Overview – A multifaceted Methodology

- The Sustained System Performance Measure (SSP) is a very effective ***Methodology*** for evaluating and predicting performance on difference computational systems
- SSP has been evolving over 10 years with at least three discrete stages
- SSP incorporates the sense of an evolving system as well as the typical measures of performance
- SSP can be a composite benchmark
- SSP can be used to assess and compare cost effectiveness of systems and proposals
- Assess the entire system – CPUs, Memory Subsystem, Interconnect and system software

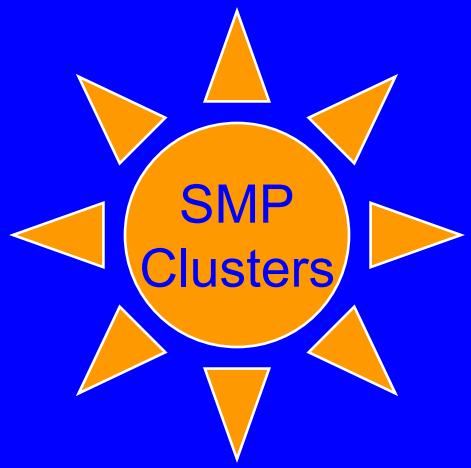
SMP Clusters Vary in Many Ways

Number of CPUs sharing single uniform memory

Memory Hierarchy
NUMA - Distributed

Memory Hierarchy for SMP
Bus – switched; Speed; Caches

CPU Performance
Clock rate;
Pipelining;
Prediction; Vectors



On-line storage
Local – global; Serial – parallel
Interconnection; Hierarchy

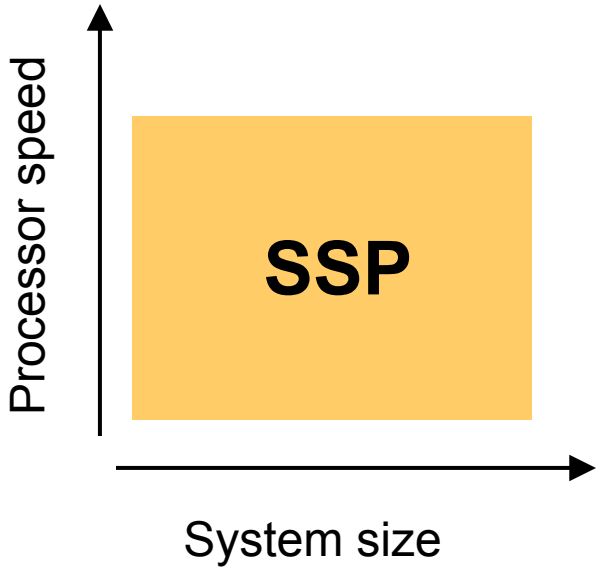
Communications within system
Topology; Speed-Bandwidth; Latency; Where connected into SMP; Protocols

System software
Single vs multiple images; Micro kernel vs full O/S on SMPs; Scalability of O/S; 32 or 64 bit; Location of Functional Layers

Vendor Proprietary UNIX derivative
Mature; HPC functions; Each implementation different
(More a concern for system support than for system users. Most NERSC codes run on multiple systems)

Open Source LINUX Derivative
Operating System Companies doing distributions (RedHat, SuSE, Caldera, TurboLinux, Scyld…);
System Vendor Companies not doing distrribution (IBM, SGI, Compaq, …);
Cluster Integrators/Consultants (HPTi, Paralogic..);
Linux Community (Extreme Linux, Cplant, Chiba City…);

SSP is a Composite Metric



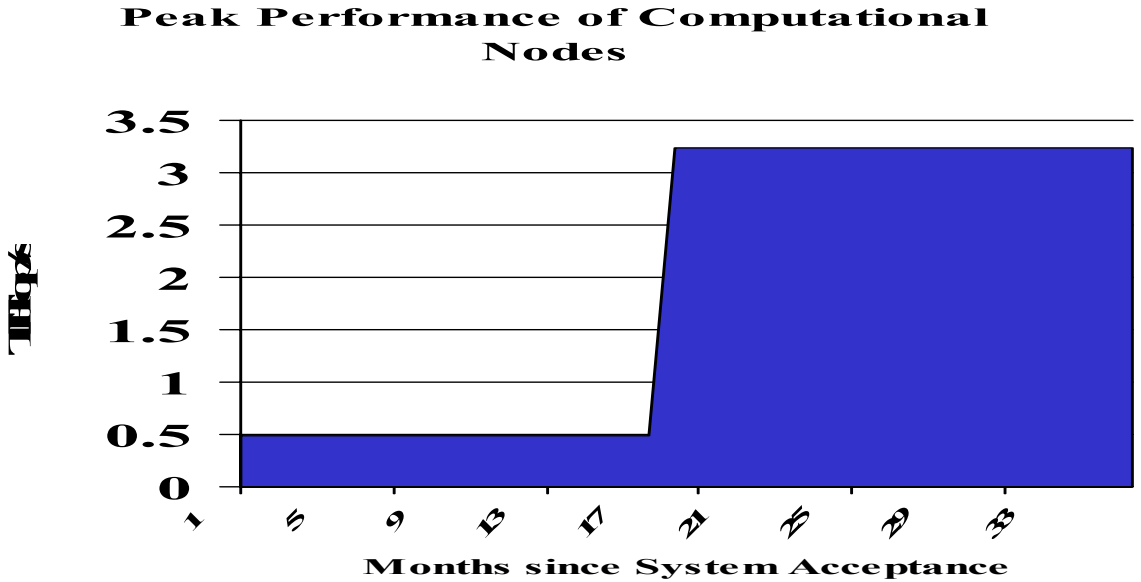
- Dependent on processor speed and system size
- Compare very different systems
- Two means for vendors to provide requested performance
- Single performance target for vendor commitment

Peak Measures Do NOT Indicate a System is Effective for Science

- Peak Operations/sec is a very misleading measure of system performance
  - Says nothing about how much performance can be applied to scientific codes
- Percent of Peak Performance achieved varies widely
  - T3E as an example
    - 644 processors at 900 Mflop/s PE = 580 Gflop/s Peak
    - SSP-1 measured 29.6 Gflop/s per month for the system
      - ~46 MFlop/s/PE
      - 5.1% of peak
    - Studies of major NERSC applications indicate system is about 67 Gflop/s
      - ~104 MFlop/s/PE
      - 11.6% of peak
    - Gordon Bell prize winning code LSMS was 256 Gflop/s
      - ~398 MFlop/s/PE
      - 44.1 % of peak

Time and System Evolution

- **The Question - How does one assess the assess the value of systems that have different performance at different times.**
- **The Solution – Integrate the the performance over a fixed time period.**
  - **We talk about SSP values as both the integrated value or the integrated value divided by the number of months in the time period.**

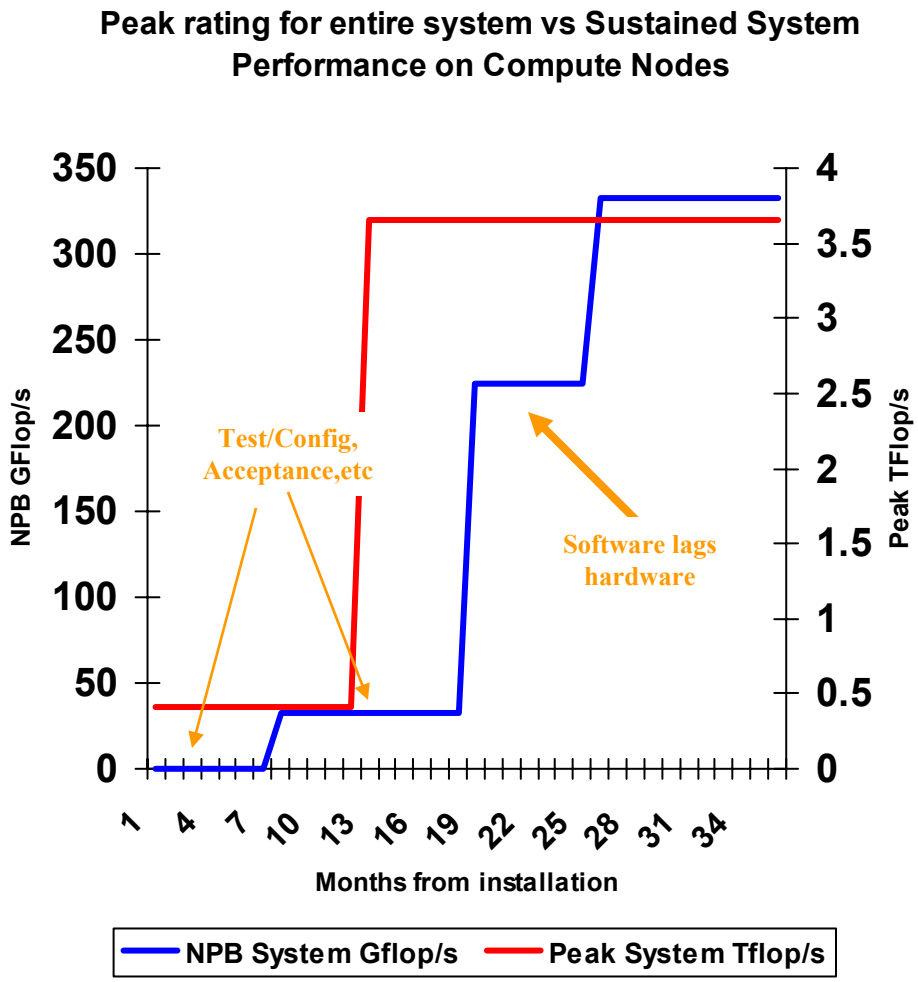


SSP Consists of Multiple Codes

- SSP is a composite measure achieved by five or six codes.
  - Full Applications
  - Pseudo Applications
  - Kernels
- Multiple codes allow different algorithms and disciplines to be represented.
  - Allows a wider evaluation
  - Provides a more robust measure
  - Able to study the performance of each code as well of overall
  - Able to simplify the measures

SSP Is Better Indicator than Peak/Linpack for the scientific value of systems

- Estimates the amount scientific computation that can really be delivered over time for a system of a constant effectiveness
  - Peak performance is misleading
  - Indicate the lower level of what a good code should get
- Motivated earlier delivery of technology
  - but only when it can be measured and is usable by scientific codes

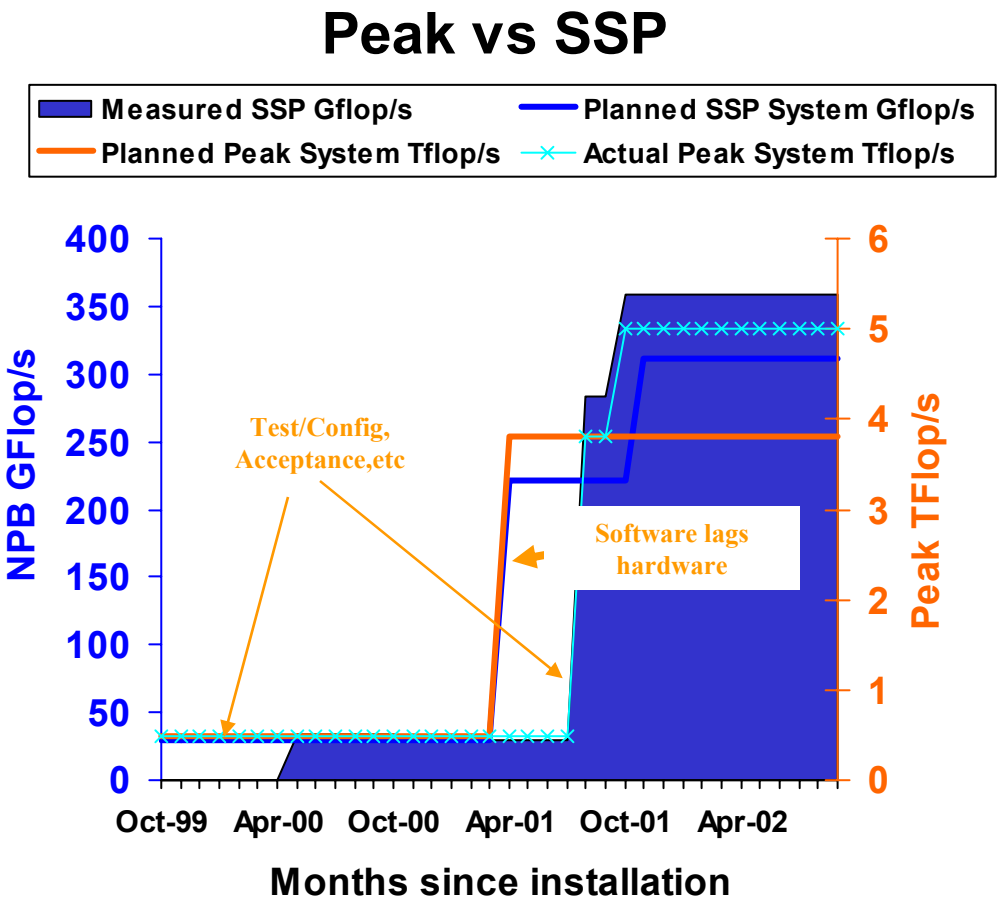


SSP Version 1

- Used for NERSC-3 procurement
- Used the six Floating Point NAS Parallel Benchmarks
  - NPBs were well proven, widely used for over 6 years
  - NPBs were good representatives of the NERSC Workload
  - NPBs scaled to expected system size
    - All ran 256 tasks
- NPBs were conservative in the sense it is harder to achieve high performance than many NERSC applications
- SSP-1 (NPBs) are a tough but honest measure for vendors and typically indicates the lower level of what a good code gets on the system
  - SSP-1 indicated T3E is a 30 GFlop/s system yet Gordon Bell prize code runs at >250 GFlop/s
  - SSP-1 indicated 365 Gflop/s per month on the Phase 2 IBM SP yet several codes were running well at 1 Tflop/s and some over 2 Tflop/s
- The six codes had equal weight in determining the overall result

Sustained System Performance Results

- Estimates the amount of scientific computation that can really be delivered over a time period
  - Measures related to workload and benchmarks
  - Depends on delivery of functionality
  - The higher the last number is the better since the system remains at NERSC for more than 3 years
- NERSC focuses on the area under the measured curve



SSP = Measured Performance * Time

SSP Version 2

- Used for NERSC-4 procurement
- Used five full Application codes
- Selection Criteria included
  - Having a distribution of disciplines and algorithms
  - Scalability – both codes and data sets
  - Portability
  - Able to be instrumented
  - Able to distribute
- NERSC polled the user community and asked for volunteer codes
- The five codes have equal weight in determining the overall value

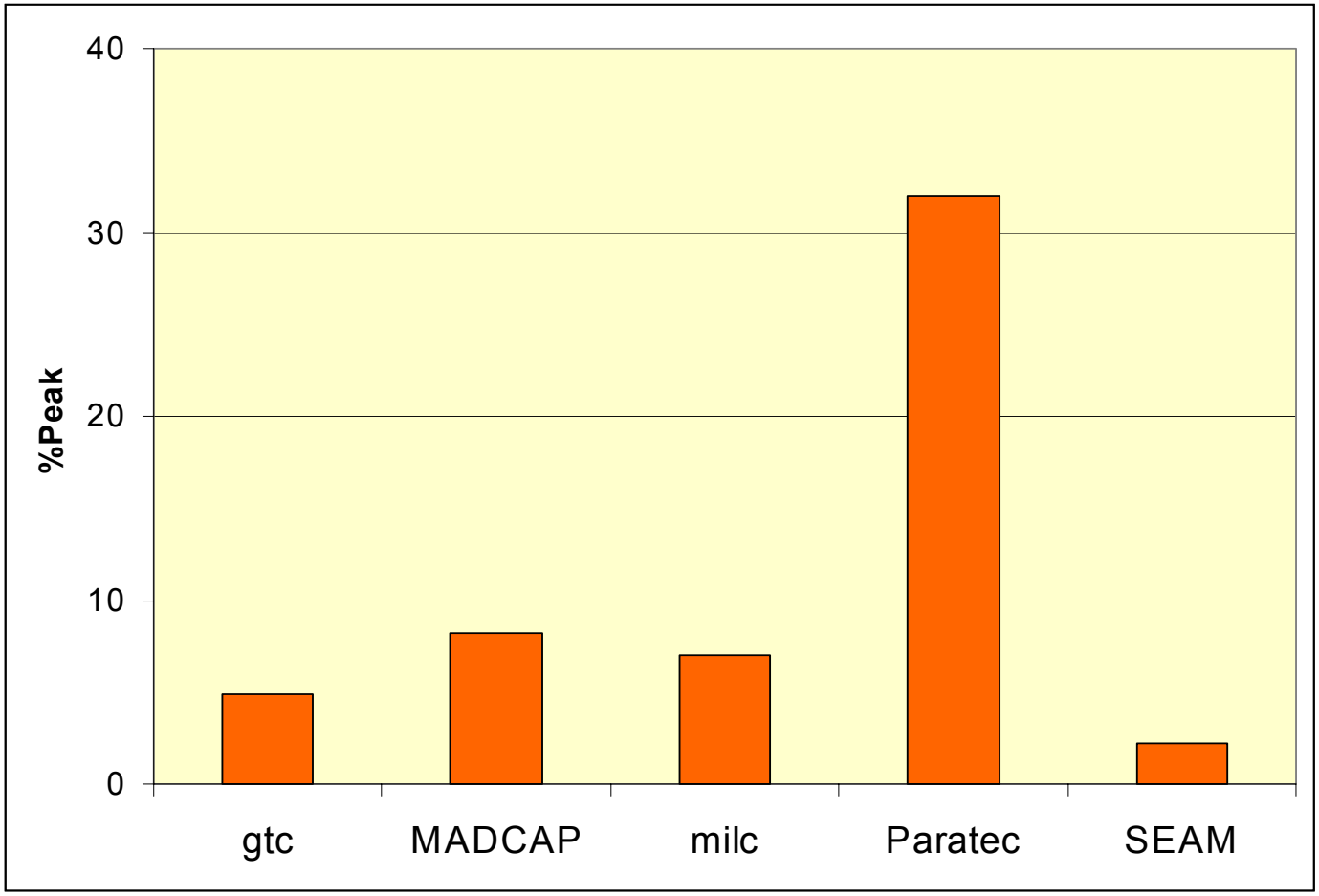
SSP-2 Component Applications

Application	Scientific Discipline	Algorithm or Method	**MPI Tasks**	**System Size**	**Seaborg Timing (sec)**	**GFlop/s**	**Percent of Peak**
gtc*	Plasma Physics – (SciDAC)	Particle-in-cell	**256**	**10^7 ions**	**1682**	**19.4**	**5.1**
MADCAP*	Cosmology (SciDAC)	Matrix inversion	**484**	**40000x 40000**	**903**	**60.7**	**8.4**
Milc*	Particle Physics (SciDAC)	Lattice QCD	**512**	**32^3x64**	**1031**	**40.7**	**5.9**
NAMD	Biophysics	Molecular dynamics	**1024**	**92224 atoms**	**379**	**31.0**	**2.0**
NWChem	Chemistry	Density functional	**256**	**125 atoms**	**2367**	**2.52**	**0.7**
Paratec*	Material Science	Density functional	**128**	**432 atoms**	**1386**	**59.0**	**30.7**
SEAM*	Climate (SciDAC)	Finite element	**1024**	**30 days**	**494**	**109.0**	**7.1**

SSP-2 Application Characteristics

	I/O	dense	sparse	all2all	scalability
gtc	low	medium	high	medium	medium
MADCAP	high	high	low	medium	medium
milc	medium	high	low	medium	medium
Paratec	medium	high	low	high	low
SEAM	low	medium	low	medium	high

SSP-2 Application Performance



SSP Does Make a Difference From the NERSC-4 Procurement

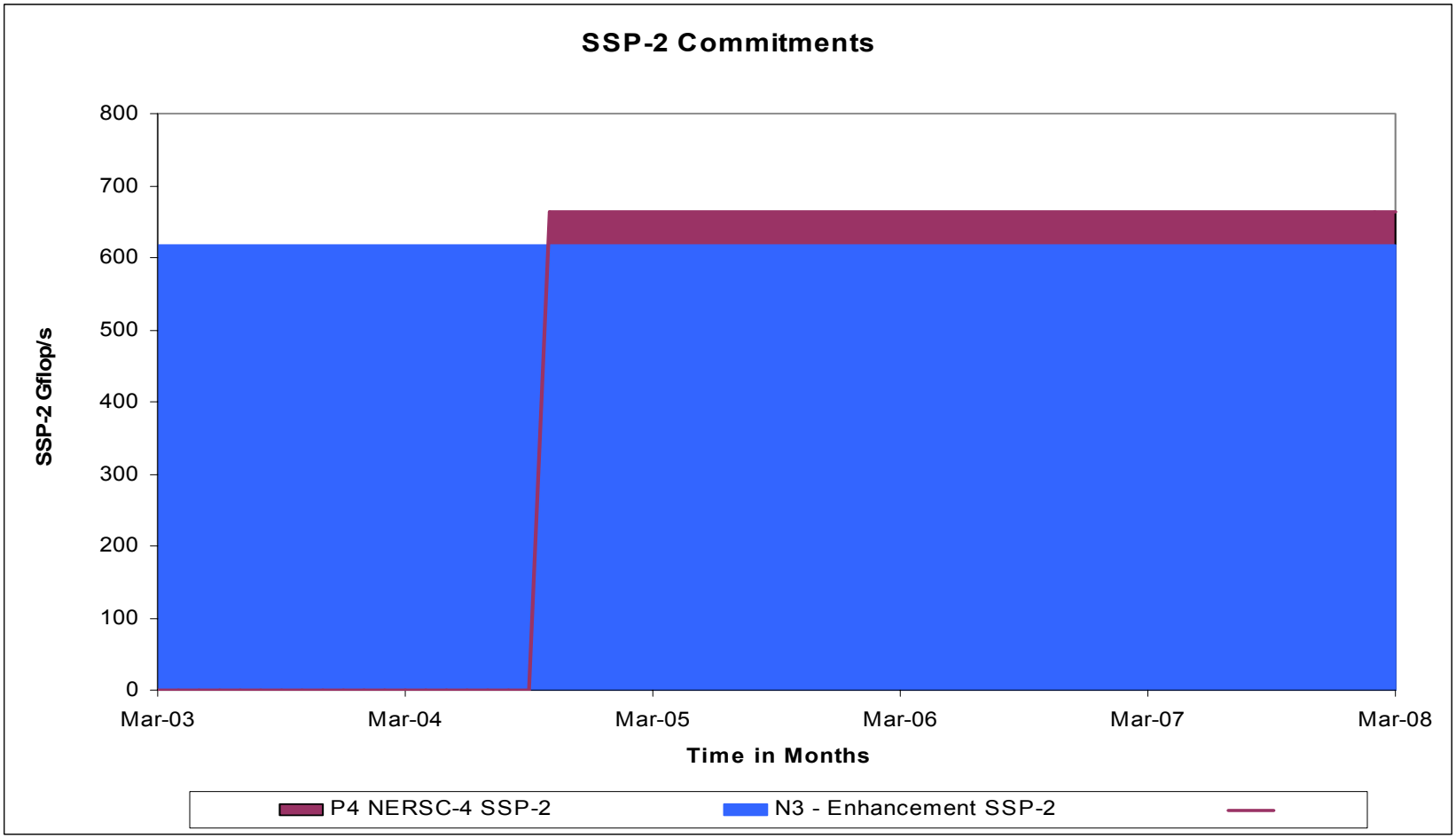
- **IBM offered a Power 4 solution that offered**

  663 Gflop/s on NERSC SSP

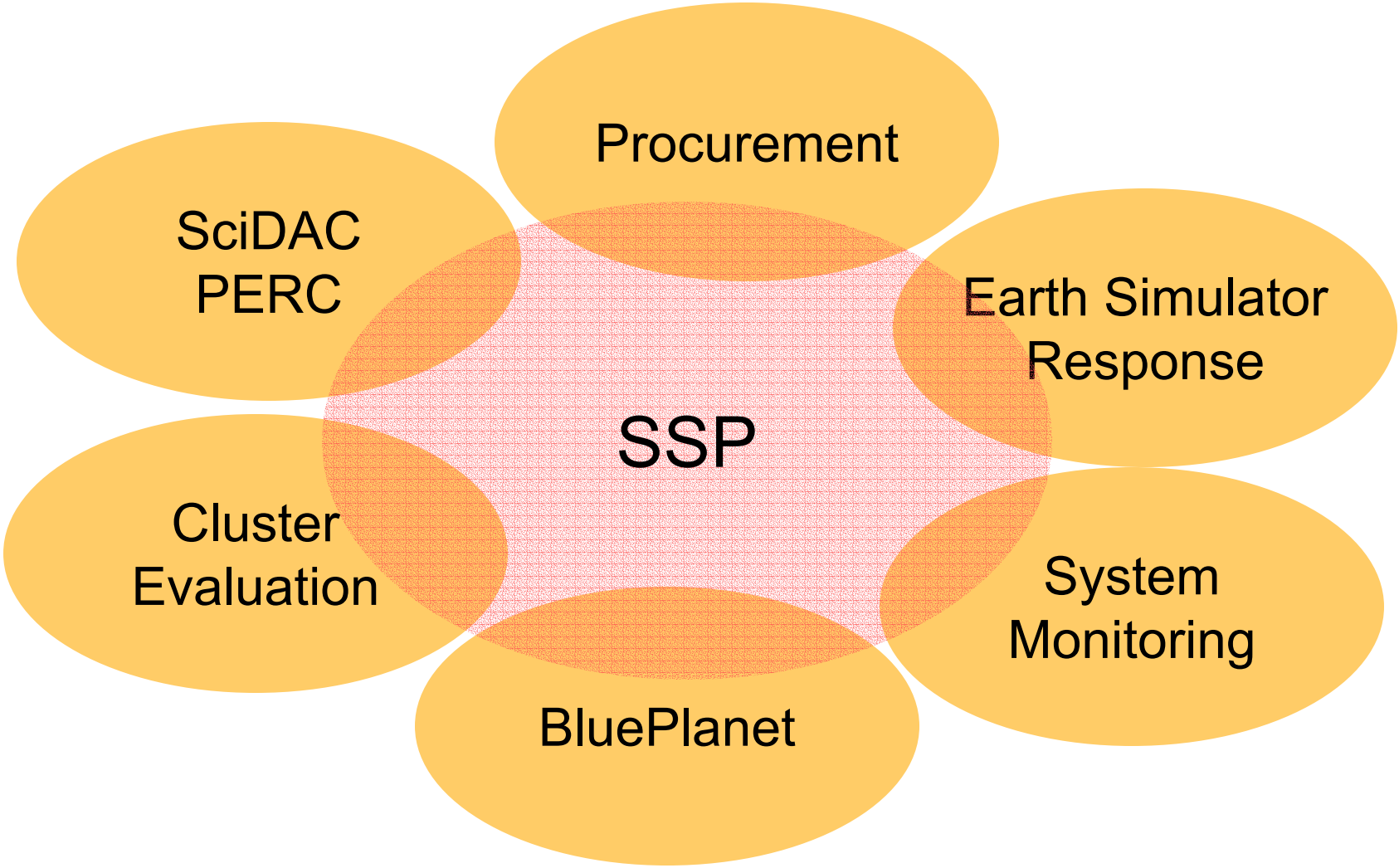
  Within our budget constraints
- **The NERSC – 3 base system delivers**

  618 Gflop/s on NERSC SSP
- **IBM offered a Power 4 system that had**
  - **Same base cost as NERSC-3**
  - **Was available to NERSC users only in mid to late 2004**
  - **Offered only a 7% performance improvement 3 years after NERSC-3**

Sustained System Performance



NERSC Activities involving the SSP Methodology



Summary

- SSP is a time proven measure for both understanding performance of current and future system
- It provides an excellent test method for selecting systems for purchase
- It is representative of the NERSC workload and the performance it would have on the systems
- It is an indicator of the computational value for DOE Science
  - And an indicator of the number of MPP hours NERSC provides