GENERAL DESCRIPTION OF THE SYSTEM

Mandatory Requirements [Required]

M1. A balanced, multi-user, Unix-based computer system supporting batch and interactive use is required. The system must accomplish the scientific computing needs of the EMSL community described in *The Users' Requirements Document*. In addition, the final system must be 64-bit and easy to use and administer.
M2. The final component of a staged install must be delivered no later than one calendar year after installation of the initial components.
M3. Nearly all available resources of the system must be usable by a single parallel job.
M4. A small test system representative of the large final installation is required.

Value Related Factors [Desired]

V1. An experimental computing system to test the offeror's most advanced hardware and software capabilities is desired.

PHYSICAL

Mandatory Requirements [Required]

It is the offerors responsibility to propose a system that will fit into the MSCF. An AutoCAD V14 drawing of the MSCF can be downloaded from the PNNL web site at: at http://www.emsl.pnl.gov/auth/mscf/hpcs2/ . The offeror is strongly encouraged to have a site-planning representative visit the MSCF. As of April, 2001, Battelle has requested funding to provide the following limits:

M5. Facilities for the proposed system cannot exceed 2000 ft^2 including service clearance;
M6. Power cannot exceed 750KVA;
M7. Floor load cannot exceed 350lbs. per square foot and 1000lb point load;
M8. Cooling requirements cannot exceed 220 tons of AC.

Value Related Factors [Desired]

V2. Power does not exceed 380KVA; and,
V3. Cooling requirements does not exceed 107 tons of AC.

Note: If funding does not materialize to modify the facilities of the MSCF, the value related factors V2 and V3 shall revert to Mandatory Requirements.

PARALLEL APPLICATION PROCESSOR SPECIFICATIONS

Mandatory Requirements [Required]

M9. Must conform to IEEE-754 64-bit arithmetic for 64-bit integer, 64-bit floating point. Further, 64-bit pointer and 64-bit file pointer hardware or software support are required.

Value Related Factors [Desired]

V4. Thirty-two bit compute model;

V5. Extended precision (> 64-bit) hardware and software support for floating point, arithmetic, pointer, and file pointers;
V6. Virtual memory to provide applications with total usable memory at least twice the physical memory per CPU.

INTERPROCESSOR COMMUNICATIONS

Mandatory Requirements [Required]

M10. Must provide MPI and efficiently support the MPI+GA programming model. Must also be able to allocate and use a single global array that includes at least 50% of the total memory in the system (i.e., at least 0.5 Tbyte).

Value Related Factors [Desired]

V7. MPI-2, optimized GA, and parallel I/O libraries;
V8. H/W support for large SMP/NUMA configurations that give very low latency (<1μs) for remote memory reference (GA-get).

AGGREGATE SYSTEM PERFORMANCE AND CAPACITY

Mandatory Requirements [Required]

M11. The minimum aggregate peak 64-bit IEEE floating-point speed is 4 Tflop/s (as measured by the aggregate best single-processor DGEMM that includes degradation inside SMP nodes).
M12. The minimum aggregate memory for a 4 Tflop/s system is 1 Tbyte. Suitable memory for a 10 Tflop/s system must be at least 2 Tbytes.
M13. There must be at least 1 Gbyte of physical memory per processor for use by an application, plus whatever memory is required by the operating system and related processes, including I/O, network, and communication buffers.
M14. At least 25% of the compute processors must have scalable, high bandwidth I/O to external disk with at least 25 (Mbyte/s)/(Gflop/s) (i.e., 50 Mbyte/s on a 2 Gflop/s processor) to an aggregate of at least 50 Tbytes of data. [Since these applications can exploit non-shared file systems, Battelle has historically attached disks locally to each processor because this has given the best price performance. However, other solutions are acceptable].
M15. A shared, high-performance parallel file system that is accessible from all nodes in the system with a capacity of at least 50 Tbytes is required. A job running on all nodes in the system must be able to write out the entire contents of memory in 10 minutes, which translates into a minimum speed of 1.7 Gbyte/s for 1 Tbyte of memory.
M16. A shared file system of at least 2Tbytes for user files is required.
M17. Gigabit Ethernet network connections providing 3.2-gigabits/second aggregate bandwidth and supporting ether-channel link aggregation for redundancy and throughput enhancement.

Value Related Factors [Desired]

V9. Extra disk, memory, processing power;
V10. Low latency for random access IO on shared-file systems.

DEVELOPMENT SOFTWARE

Mandatory Requirements [Required]

M18. Basic support for Fortran77, Fortran90, C, and C++ is required. In addition, preprocessors, compilers, and associated libraries must provide inter-language interoperability. Compilers must support 64-bit integer, floating point, and pointer data types with compiler options. Cray-style pointers must be a supported data type.

M19. Interactive and postmortem parallel debugging environment must support:

a. source level;
b. mixed language;
c. mixed multi-threaded/MPI applications.

M20. Source level sequential and parallel CPU time profiling are required.

M21. Python, Perl, and Java are required.

Value Related Factors [Desired]

V11. Fortran95; and,

V12. Programming and interoperability of Java and/or Python with the basic compiler languages like Fortran, C, and C++.

Libraries

Mandatory Requirements [Required]

M22. Optimized 64-bit BLAS1,2,3 and LAPACK numerical libraries;

M23. Support of POSIX conforming threaded programming model within SMP nodes;

M24. X11 libraries available on all nodes.

Value Related Factors [Desired]

V13. Optimized multiprocessor distributed-memory libraries (e.g., parallel scilib, SCALAPACK, parallel BLAS, parallel FFT, …);

V14. High performance parallel I/O (e.g. MPI-I/O);

V15. High resolution timer;

V16. Hardware performance counters for CPU, such as flops, cache performance, etc are desirable with software access to get the information (e.g., API or compile/run time options), to organize the data for parallel applications, and to analyze the results ideally in the context of the source code.

SYSTEM MANAGEMENT HARDWARE AND SOFTWARE

Mandatory Requirements [Required]

M25. Scalable cluster management tools;

M26. Complete resource management and job accounting solution;

M27. Network installation, remote power management, remote console access;

M28. Hardware and software monitoring;

M29. Configuration management (for system files, software packages and patches);

M30. Unix process-level accounting;

M31. Support for AFS;

M32. Offeror must provide a security solution that includes support for:

a. Kerberos 5 authentication;
b. Ability to display a login banner;
c. Ability to audit successful and attempted connections;
d. Secure remote execution and file transfer.

Value Related Factors [Desired]

V17. Onsite personnel;
V18. Checkpoint/Restart, Gang Scheduling;
V19. Scheduler support for reservations, backfill, checkpoint/restart, metascheduling;
V20. Event management;
V21. Performance monitoring;
V22. The global filesystem should be HSM compatible (e.g. DMAPI compatibility);
V23. Buildable source code for OS and delivered software;
V24. Tools for analyzing kernel crash/dump and problem diagnostics; and,
V25. Site-integrated security mechanism for single-logon authentication that utilizes strong authentication (ability to run with no clear text passwords and no .rhosts)

DOCUMENTATION AND TRAINING

Mandatory Requirements [Required]

M33. Full documentation including online copies, paper copies (additional copies upon request), and a limited right to copy documentation.
M34. A minimum of 1000 person-hours of on-site training.

Value Related Factors [Desired]

V26. Web based documentation.

AVAILABILITY

Mandatory Requirements [Required]

M35. Must have three-year support and warrantee including [24 x 7] 4-hour onsite hardware support and access to knowledgeable software support during working hours (PST).
M36. The system must be able to demonstrate 99% system availability during a 30-day acceptance test and an average 98.5% availability over each year of the contract.

Value Related Factors [Desired]

V27. Automatic hardware / software diagnostics on first failure;
V28. Enhanced support structure including online support and patch download, 24X7 software support, and onsite hardware spare parts;
V29. Ability to perform corrective or preventative software and hardware maintenance on a portion of the system without impacting the rest of the system;
V30. Low life-cycle costs (maintenance, operation, integration, conversion); and,
V31. Reliability of service:

a. High overall and scheduled availability;

b. High mean time between interrupts;
c. Low mean time to restore;
d. High mean time between system failures;
e. Low time required to perform preventative maintenance.

[the next page begins the Application Benchmarks Section]

APPLICATION BENCHMARKS

The offeror is expected to complete the mandatory sequential and parallel benchmark tests described below. The benchmark codes, data files and README files can be downloaded from the PNNL web site at http://www.emsl.pnl.gov/auth/mscf/hpcs2/.

Benchmark compilation and execution requirements:

A. The benchmark tests are to be run on a system that is as close as possible to the configuration of the proposed system.
B. The compilers, operating system, and any special system features and/or tools used in the benchmarks must be part of the offered supported software configuration. The offeror can run the benchmark in full 64-bit or with 32-bit integers, pointers, and libraries (including MPI). 64-bit benchmark results are considered a value related factor for the proposal response. When the benchmarks are run in 32-bit, a projection of the benchmark performance on the proposed 64-bit platform is required. The offeror will be required to run the mandatory benchmarks in 64-bit mode as part of the acceptance tests.
C. The offeror must describe the hardware and software used for the benchmarks. In addition, the difference between the benchmark system and proposed system must be described.
D. For the sequential benchmarks 1.A. – 1.G. and the parallel benchmarks 2.A. – 2.C, the offeror must project the measured results from the benchmark system to the proposed system, and the projection method must be explained and justified in the proposal. For the parallel benchmarks 2.D. – 2.G., only a projection of the measured numbers on the benchmark system to the same number of processors on the proposed system is required.
E. The offeror must provide the standard output results produced by the benchmark test. Offeror must document code changes needed (when applicable), and compiler scripts (Makefiles) or compiler and compile options used to achieve the results for the sequential benchmarks 1.A. – 1.F. and parallel benchmarks 2.B. – 2.H. For 1.G. and 2.A. only the benchmark results are required.
F. For sequential benchmarks 1.A. – 1.E. and parallel benchmark 2.C., no changes to the source code are allowed unless explicitly stated in the benchmark descriptions listed below. For the application benchmark codes NWChem, NWGrid / NWPhys, and Lattice-Boltzman no algorithmic changes are permitted but it is permissible to modify the source for porting or correctness, and also to use tuned versions of underlying libraries, e.g., ARMCI, MPI, BLAS, etc.
G. All mandatory benchmarks must be run separately on a dedicated system. No other jobs may be run on the system during the benchmark test.

1. Mandatory sequential benchmarks:

A. LINPACK 100 Benchmark: This benchmark must be run sequentially, using double precision arithmetic.
B. LINPACK 1000 (TPP) Benchmark: This benchmark must be run sequentially, using double precision arithmetic. The standard LINPACK TPP rules for code changes and running the benchmark must be followed.

C. Spec CPU2000 Benchmarks: Only the combined single-value results, using the non-baseline (aggressive optimization) sequential SPECint2000 and SPECfp2000 version are required. PNNL is not providing these benchmark tests and offerors who do not have the Spec CPU2000 version 1.1 can obtain them at http://www.specbench.org/ .
D. Livermore Fortran Kernels (LFK) Benchmark: Only the sequential Standard Benchmark LFK Test is required using the run rules as outlined in the LFK README file.
E. STREAM Benchmark: Changes to the second_cpu.f source code are allowed if a different timer is needed.
F. NWChem Sequential Benchmark: The following benchmark tests must be run to completion on a single processor:
    a. UO2_DFT
    b. BR2_DK
    c. PSPW
G. DGEMM Performance Benchmark: The sequential DGEMM performance of source code or a library provided by the offeror, and delivered with the proposed system, is required. The benchmark must be run on a single processor and, if an SMP solution is proposed, on all processors in an SMP node (running a copy of DGEMM on each processor in the SMP node). The offeror must provide the performance numbers in Flops/sec and the problem size used to achieve this performance.

2. Mandatory parallel benchmarks:

A. MPI latency: Measured MPI 1/2 ping-pong latency. If an SMP cluster solution is proposed, the MPI latency between processors inside an SMP and between processors of different SMP nodes is required.
B. Massively Parallel LINPACK Benchmark: The measured R_{max} (maximum LINPACK performance achieved), N_{max} (problem size for achieving R_{max}), and $N_{1/2}$ (Problem size for achieving half of R_{max}) are required. A projection of these numbers for nearly all compute processors on the final proposed system is required.
C. NAS Parallel Benchmark: Only the Class C MPI results in NAS version 2.3 are required. Only the results from the 0 % changes run rule version are required. The objective is to minimize the wall clock time in each of the benchmark cases. The offeror must provide the number of processors and the wall clock time for which the minimum wall clock time is achieved.
D. ARMCI Benchmark: We want measured the best- and worst-case latency and bandwidth for remote memory read using the ARMCI library which measures the time taken by process zero (0) to access data "owned" by the other processes. The offeror must describe process placement relative to the memory hierarchy or communication topology for each reported number. The specific benchmark requirements will depend on the proposed solution. If a single processor per node solution is proposed the benchmark must be run on 4 processors. When an SMP cluster solution is proposed, two (2) different benchmarks are required:
    a. Communication between SMP nodes. The benchmark must be run using 4 SMP nodes and only 1 processor per SMP; and,
    b. Communication within an SMP node. The benchmark must be run using at least 2 processors in an SMP node.

E. NWChem Parallel Benchmark: The following benchmark tests must be run to completion:
   a. DFT_SiSO3: The benchmark test must be run on 1, 2, 4, 8, 16, 32 and 64 processors.
   b. MP2_(H2O)7: The benchmark test must be run on 1, 2, 4, 9, 18, 36 and 63 processors.
   c. MD_HAD: The benchmark test must be run on 27 and 54 processors.
F. NWGrid / NWPhys Benchmark: The following benchmark tests must be run to completion:
   a. SPEEDUP: The benchmark must be run on 8, 16, 32, and 64 processors.
   b. SCALEUP: The benchmark must be run on 32 and 64 processors.
G. Lattice-Boltzman (LB) Benchmark: The following benchmark tests must be run to completion:
   a. SMALL: The benchmark must be run on 8 processors.
   b. LARGE: The benchmark must be run on 64 processors.
H. Throughput Application Benchmark (TAB): The offeror is required to demonstrate the ability to run, simultaneously and efficiently, multiple programs in the batch environment to be included in the proposed system. The benchmark consists of a single run of multiple copies of the benchmarks 2.E. – 2.G. on a benchmark system of 64 processors. The benchmark runs to be used in this benchmark are listed in the table below. The runs must be submitted to the batch system in the order they appear in the table, and 2-9 must be submitted after run 1 has started its execution. Start time, end time, and elapsed time for each individual run, as well as for the whole Suite of runs (1-9) must be reported.

#	Benchmark run	Name	# of processors
1	2.E.c.	NWCHEM: MD_HAD	54
2	2.E.b.	NWCHEM: MP2_(H2O)7	36
3	2.F.b.	NWGrid / NWPhys: SCALEUP	32
4	2.E.a.	NWCHEM: DFT_SiSO3	16
5	2.G.a.	Lattice-Boltzman: SMALL	8
6	2.F.a.	NWGrid / NWPhys: SPEEDUP	16
7	2.E.c.	NWCHEM: MD_HAD	27
8	2.E.b.	NWCHEM: MP2_(H2O)7	9
9	2.G.a.	Lattice-Boltzman: SMALL	8

3. Value Related Benchmarks and Value Related Factors:

A. NWChem Value related Benchmark:
   a. MP2_(H2O)9: The benchmark should be run on at least three (3) of the following number of processors: 126, 252, 504, 1008.
   b. MD_HAD2: The benchmark test should be run on 216 processors.
B. NWGrid / NWPhys Value related Benchmark: The benchmark SCALEUP from the parallel benchmark 2.F.b. should be run on at least three (3) of the following number of processors: 128, 256, 512, 1024.
C. The proposed platform should be able to run at least the following third party codes: GAUSSIAN, ADF, MOLPRO, GAMESS, STAR-CD, CCM and MM5.

Tables to be filled with the results from the Mandatory Sequential and Parallel Benchmarks.

Table 1.: Results from the Mandatory Sequential Benchmarks. The results must be taken from the output produced, except for 1.G.

Benchmark		Parameter	Unit	Measured on benchmark system	Projected for proposed system
1.A.	LINPACK 100	Flop Rate	Mflops/s		
1.B.	LINPACK 1000	Flop Rate	Mflops/s		
1.C.	Spec CPU2000	Geometric Mean	---		
1.D.	LFK	Geometric Mean Maximum Rate Minimum Rate	Mflops/s Mflops/s Mflops/s		
1.E.	STREAM: Function Copy Function Scale Function Add Function Triad	 Rate Rate Rate Rate	 Mbyte/s Mbyte/s Mbyte/s Mbyte/s		
1.F. 1.F.a.	NWChem: UO2_DFT	 CPU time output Wall clock time output	 seconds seconds		
1.F.b.	BR2_DK	CPU time output Wall clock time output	seconds seconds		
1.F.c.	PSPW	CPU time output Wall clock time output	seconds seconds		
1.G.	DGEMM	Flop Rate Matrix dimensions used	Mflops/s ---		

Table 2.A.: Results from the Mandatory Parallel Benchmarks 2.A., 2.B., 2.C., 2.D., and 2.G. Results must be copied from the produced outputs, except for 2.A. and 2.B.

Benchmark		Parameter	Unit	Measured on benchmark system	Projected for proposed system
2.A.	MPI latency	Latency	μs		
2.B.	MP LINPACK	R_{max} N_{max} $N_{1/2}$	Mflops/s --- ---		
2.C.	NAS: BT	# of processors Run time	--- seconds		
	CG	# of processors Run time	--- seconds		
	EP	# of processors Run time	--- seconds		
	FT	# of processors Run time	--- seconds		
	IS	# of processors Run time	--- seconds		
	LU	# of processors Run time	--- seconds		
	MG	# of processors Run time	--- seconds		
	SP	# of processors Run time	--- seconds		
2.D. 2.D.a.	ARMCI: Between (SMP) nodes	Get Latency Get Bandwidth Acc Latency Acc Bandwidth	μs Mbyte/s μs Mbyte/s		
2.D.b.	Inside SMP node	Get Latency Get Bandwidth Acc Latency Acc Bandwidth	μs Mbyte/s μs Mbyte/s		
2.G. 2.G.a.	Lattice-Boltzman: SMALL	Total run time from output	seconds		
2.G.b.	LARGE	Total run time from output	seconds		

Table 2.B.: Results from the Mandatory Parallel Benchmarks 2.E. All results are in seconds and must be copied from the NWChem output.

Benchmark NWChem		Number of processors	Measured on benchmark system		Projected for proposed system	
			CPU time	Wall clock time	CPU time	Wall clock time
2.E.a.	DFT_SiSO3	1 2 4 8 16 32 64				
2.E.b.	MP2_(H2O)7	1 2 4 9 18 36 63				
2.E.c.	MD_HAD	27 54				

Table 2.C.: Results from the Mandatory Parallel Benchmarks 2.F. NWGrid / NWPhys. All results are in seconds and must be copied from the NWGrid (NWGrid setup time, Total time on all proc, and Minimum and Maximum on proc) and NWPhys (NWPhys Grind Time, Total time on all proc, and Minimum and Maximum on proc) output.

Benchmark NWGrid / NWPhys	Number of processors	Module / code	Measured on benchmark system			Projected for proposed system		
		Min	Total	Min	Max	Total	Max	Max
2.F.a. SPEEDUP	8	NWGrid NWPhys						
	16	NWGrid NWPhys						
	32	NWGrid NWPhys						
	64	NWGrid NWPhys						
2.F.b. SCALEUP	32	NWGrid NWPhys						
	64	NWGrid NWPhys						

Table 2.D.: Results from the Mandatory Parallel Benchmarks 2.H. Throughput Application Benchmark.

Batch system used by offeror:					
#	Benchmark	# of processors	Start time (hh:mm:ss)	End time (hh:mm:ss)	Elapsed time (seconds)
1	2.F.b.	32			
2	2.E.b.	36			
3	2.E.c.	54			
4	2.E.a.	16			
5	2.G.a.	8			
6	2.F.a.	16			
7	2.E.c.	27			
8	2.E.b.	9			
9	2.G.a.	8			
Total Suite from benchmark 2.H.					

<u>Tables to be filled with the results from the Value related Benchmarks and Value related Factors.</u>

Table 3.A.: Results from the Value related Benchmarks 3.A. All results are in seconds and should be copied from the NWChem output.

Benchmark NWChem		Number of processors	Measured on benchmark system		Projected for proposed system	
			CPU time	Wall clock time	CPU time	Wall clock time
3.A.a.	MP2_(H2O)9	126 252 504 1008				
3.A.b.	MD_HAD2	216				

Table 3.B.: Results from the Value related Benchmarks 3.B. NWGrid / NWPhys. All results are in seconds and must be copied from the NWGrid (NWGrid setup time, Total time on all proc, and Minimum and Maximum on proc) and NWPhys (NWPhys Grind Time, Total time on all proc, and Minimum and Maximum on proc) output.

Benchmark NWGrid / NWPhys	Number of processors	Module / code	Measured on benchmark system			Projected for proposed system		
		Min	Total	Min	Max	Total	Max	Max
3.B.: benchmark 2.F.b. SPEEDUP	128	NWGrid NWPhys						
	256	NWGrid NWPhys						
	512	NWGrid NWPhys						
	1024	NWGrid NWPhys						

Table 3.C.: Results for Value related Factor 3.C., third party codes that can be run on the offeror's platform.

Software	Will run on offerer's platform (yes / no)
GAUSSIAN	
ADF	
MOLPRO	
GAMESS	
STAR-CD	
CCM	
MM5	