Leistungsbeschreibung „HPC-Cluster“

Weitgehend homogene Architektur, CPU-/GPU-Typ, homogene System-, Entwicklungs- und Management-Umgebung, Batchsystem, paralleles Filesystem

Für die Realisierung aller folgenden Positionen stehen 2,15 Mio. € inkl. MwSt. zur Verfügung. Eine hohe Wertung erhalten gleichrangig:

1.) eine hohe aggregierte Performance des Gesamtsystems: Cluster-TFLOPS; Core-/Node-Memory; paralleles Filesystem und I/O (dabei werden die Benchmarks höher gewertet als theoretische Werte wie Peak-TFLOPS),

2.) ein effizienter Stromverbrauch bei garantierter Einhaltung der Obergrenze von 200 KW maximaler Energieaufnahme. Eine Überschreitung führt zum Ausschluss. Stromeffizientere Lösungen erhalten bei ansonsten vergleichbaren Leistungen eine höhere Wertung.

3.) ein überzeugendes transparentes Gesamtkonzept hinsichtlich sowohl aller Anforderungen der Leistungsbeschreibung inklusive Service und Support, als auch das System perspektivisch mit mindestens 50 % zusätzlichen Nodes und etwa 100 % zusätzlicher Diskkapazität möglichst reibungslos aufzurüsten zu können.

Datum Firma Gesamtpreis des unten beschriebenen Angebots (inkl. MWSt.)

0. Erläuterungen zu den Tabellen

Die erste Spalte enthält die Zeilennummer dieser Tabelle.

Die zweite Spalte beschreibt die Anforderungen, die an den Anbieter gestellt werden und soll von diesem nicht verändert werden.

Die dritte Spalte ist vom Anbieter auszufüllen. Sie kann Anmerkungen in grauer Farbe enthalten, die vom Anbieter gelöscht werden können.

Die vierte Spalte enthält die Symbole „A“, „B“, „b“ oder „I“.

- „A“ steht hierbei für eine Forderung, die vom Anbieter unbedingt zu erfüllen ist.

  Angebote, die eine dieser Forderungen nicht erfüllen, werden vom Auswahlverfahren ausgeschlossen.
- „B“ ist ein Bewertungskriterium. Bewertungskriterien werden vom Antragsteller dazu verwendet, Angebote einander gegenüberzustellen.
- „b“ ist ein Bewertungskriterium mit verminderter Priorität.
- „I“ steht für ein Informationskriterium. Diese Kriterien werden nicht zur Bewertung herangezogen.

Sofern in der Tabelle nicht ausreichend darstellbar, sind zusätzliche Anlagen erlaubt. Dabei bitten wir um eine kompakte, übersichtliche Darstellung (Tabellen, Skizzen, …) und um unbedingte Einhaltung der Gliederung in diesem Dokument.

Unvollständige Angebote werden nicht gewertet. Ein vollständiges Angebot versteht sich inklusive einer Beantwortung aller Tabellenzeilen, auch solcher in Kategorie „I“.

1. Compute- und Login-Nodes

#	Forderung	Angebot	Kat.
1	**Anzahl der Compute-Nodes, Anzahl der Cores im gesamten Cluster**		I
2	Node (Typ/Bezeichnung): Anzahl Sockets pro Node, Anzahl CPUs pro Node, Anzahl Cores pro CPU, Anzahl Cores pro Node		B
3	Mindestens 2 Steckplätze für eventuelle spätere Aufrüstung mit HDD/SSD		A
4	Angaben zu verfügbaren Steckplätzen und Typ für eventuelle spätere Aufrüstung mit HDD/SSD (z.B. 2 * 2,5“ HDD oder SSD pro Node)		b
5	genaue CPU-Bezeichnung, Typ, Taktrate		b
6	Anzahl FPUs pro Core		I
7	Anzahl FLOPS pro Core und Takt		B
8	Anzahl Threads pro Core		B
9	Cache-Stufen und -Größen pro Core: L1 = … MB; L2 = … MB; L3 = … MB, … (evtl. gemeinsame Caches spezifizieren)		B
10	TFLOPS pro Node (Angabe: Anzahl Cores * Floating-Point-Operationen pro Core und Takt * Taktrate)		I
11	TFLOPS des gesamten Clusters (Angabe: TFlops pro Node * Anzahl Nodes)		B
12	Anzahl Nodes mit folgenden Memory-Werten: • 65 % mit 128 GB/Node • 35 % mit 256 GB/Node • 2 Nodes mit 1 TB, davon eines aufrüstbar auf 2 TB unter Weiterverwendung der bereits vorhandenen DIMMs	Angaben zu den verwendeten DIMMs, sowie zu freien DIMM-Steckplätzen und deren Nutzbarkeit für spätere Aufrüstungen sind erbeten. Die Aufrüstung des 1 TB-Nodes ist *nicht* Gegenstand dieser Ausschreibung.	B
13	**Zwei Login-Nodes typgleich bei CPU und Memory zu den Compute-Nodes** (bei verschiedenen Typen typgleich zu den leistungsstärkeren hinsichtlich „Ansprechverhalten“) als Login- und Entwicklungs-/Test-Node (Login, Dialog,		A

#	Forderung	Angebot	Kat.
	Kompilieren, kurze Jobtests)		
14	256 GB Memory auf Login-Nodes, aufrüstbar auf 512 GB unter Weiterverwendung der DIMMs		A
15	Login-Node: Core-Typ und Anzahl, Takt, Memory, Systemdisk: Typ, Anzahl, Größe, RPM, HW-Mirror		B
16	10 GbE-Port pro Login-Node (zum Uni-LAN)		A
17	**12 GPGPU-Nodes mit je 4 NVIDIA GeForce GT 980 und 256 GB RAM**	Bitte nehmen Sie diese Nodes aus dem Pool der 256 GB-Nodes von Zeile 1.12.	A
18	GPU-Node spezifizieren: • Typ GPU, CUDA-Recheneinheiten, … • Typ Host, Cores, Taktrate, CPU-Sockets, …; • Kommunikation • Leistungsbedarf für den kompletten Node		B
19	Platzbedarf im Rack im Vergleich zu CPU-Nodes		I
20	Bitte skizzieren, wie spätere Aufrüstung mit weiteren CPU-Nodes (angegebene Nodes + ca. 50 % des hier angegebenen) und GPU-Nodes (angegebene Nodes + ca. 100 % des hier angebotenen) erfolgen könnte (möglichst, ohne Komponenten auszutauschen)		I

2. Kommunikation

#	Forderung	Angebot	Kat.
1	**Kommunikation im Compute-Node** (im Socket, zwischen Sockets: Architektur und Topologie)		I
2	Speicherbandbreite eines Nodes in GB/s (pro Core, pro Socket, pro Node)		B
3	**Kommunikation Compute-Node–Compute-Node, Compute-Nodes–Fileserver:** Infiniband 50 % der Nodes mit 1:2 Blockingfaktor, 50 % mit 1:8		A
4	Infiniband-Typ		B
5	Brutto/Nettobandbreite in Gbit pro IB-Port und Latenz in µs		B
6	Anzahl IB-Ports pro Fileserver		B
7	Anzahl IB-Ports pro Compute-/Login-Node		B
8	Anzahl IB-Switche für je 50 % mit: • 1:2 Blocking-Faktor (leaf/spine/core) • 1:8 Blocking-Faktor (leaf/spine/core)		B
9	Eventuell Redundanz (mit/ohne Lastausgleich)		B
10	**Kommunikation Fileserver–Disk-Array:** Anzahl und Typ Ports zwischen Fileserver und Disk-Array: Bandbreite pro Port in Gbps (brutto/netto)		B
11	Anzahl Switches, Anzahl Ports pro Switch		I
12	Kaskadierung Fileserver ↔ Storage		I
13	Redundanz (mit/ohne Lastausgleich)		B

3. Fileserver, Paralleles Filesystem, Disk-Array, lokale Disk

#	Forderung	Angebot	Kat.
1	**Disk-Array mit mindestens 800 TB Netto** (sollte funktionell mindestens mit RAID 6 vergleichbar sein)	Unter „netto“ ist hier die letztendlich nutzbare Kapazität, also ohne RAID-„Verluste“, gemeint.	A
2	Zeiten für Rebuild nach Ausfall 1. Disk, für Rebuild nach Ausfall 2. Disk im selben Array		B
3	Controller: Anzahl, Caches, Bandbreiten in GB/s, Redundanzen (falls vorhanden)		I
4	Disk-Typ, Anzahl, Größe, RPM	Bitte beachten Sie die opimale TB/Disk-Rate wegen Performance und Rebuild-Zeiten.	B
5	**Paralleles Filesystem: GPFS oder BeeGFS** (homogenes paralleles FS für gesamten Cluster wird als vorteilhaft bewertet)		A
6	/home 15 % (von mind. 800 TB Disk-Array)	Bitte geben Sie die nutzbare Kapazität an.	A
7	FS für /home spezifizieren (Server, Clients, Metadaten, …)		B
8	/scratch bzw. /work 85 % (von mind. 800 TB Disk-Array)	Bitte geben Sie die nutzbare Kapazität an.	A
9	FS für /scratch spezifizieren (Server, Clients, Metadaten, …)		B
10	Stabilitätseigenschaften (z.B. Verhalten: bei temporären Netzausfall, Ausfall eines Fileservers, …)		B
11	Spiegelung der Metadaten		A
12	Weitere vorgesehene HW- und SW-Maßnahmen zur Gewährleitung kosteneffizienter Redundanz für /home und für /scratch: (z.B. HW: Kabel, Switches, …; SW: Spiegelung der Objektdaten, …)		I
13	Systemweit konfigurierbare Quota pro Nutzer (/home, /scratch)		B
14	Dedizierte Schnittstelle des parallelen Filesystems an Tivoli Storage Manager		b

#	Forderung	Angebot	Kat.
15	Bandbreite des parallelen Filesystems (mind. 20 GB/s; siehe Abschnitt 9)		B
16	Maximale Bandbreite eines Nodes in GB/s (siehe Abschnitt 9)		B
17	Handling und Optimierungsmöglichkeiten für die Performance bei „extremen“ Situationen: • viele kleine Files (≥ 100000) • sehr große Files (ca. 10 TB) (ansonsten generell Mix aller Filegrößen)		I
18	Erweiterbarkeit des parallelen Filesystems und des Disk-Arrays (100 % als spätere Aufrüstung denkbar) ohne Austausch vorhandener Komponenten (wie Switches, …) ist gegeben bis zu ungefähr folgender Nettokapazität		B
19	Erforderliche spätere Massnahmen zur Umsetzung des in Zeile 3.18 beschriebenen Ausbaus	Die hier angegebenen Maßnahmen sind nicht Bestandteil dieser Ausschreibung.	I
20	**Mindestens 2 dedizierte Server für paralleles Filesystem** (bitte Anzahl angeben)**, oder eine Lösung mit mindestens äquivalenter Ausfallsicherheit**		A
21	Redundanzen (mit/ohne Lastausgleich): was passiert bei Ausfall eines Fileservers, ...		B
22	Anzahl, Typ, CPUs, Cores, Memory, Ports (Typ, Anzahl)		B
23	Systemdisk (Typ, Anzahl, Größe, RPM, Mirror)		B

4. Systemumgebung

#	Forderung	Angebot	Kat.
1	**Betriebssystem: Linux x86_64 für Compute-Nodes, Login-Nodes, File- und Management-Server** (identisches OS auf allen Plattformen bevorzugt)		A
2	RHEL6 oder RHEL6-Derivat mit vergleichbarem Releasezyklus und 5 Jahren Support, mindestens durch firmeninternes Know-How des Anbieters des Clusters		A
3	**Batchverarbeitung** (Erfahrungen mit Loadlevler, SGE; SLURM wäre willkommen)		A
4	dynamische Zuweisung aller Ressourcen des Clusters pro Job unabhängig vom Node (kleinste Einheit Core/Multithread-CPU)		A
5	automatische Generierung von MPI-Hostfiles, falls nötig		B
6	*Fair-share-scheduling*-Konfiguration nach Kundenvorgaben (Shares pro Zeiteinheit und pro Nutzer, Nutzergruppe, …)		A
7	Reservierung von Ressourcen		A
8	dynamisches Füllen von „Ressourcen-Lücken“		A
9	Kalenderfunktion mit Backfill (bis zum Datum sind passfähige alte und neue Jobs noch zu rechnen)		B
10	Ressourcenmanagement arbeitet integrativ mit Lizenzmanager zusammen		b
11	Anzeige, Kontrolle und Limitierung realer Memory pro Job durch Admin und User		A
12	Hard-Limit zur Unterbindung von Paging (Angabe Methoden)		A
13	Process binding (spezifizieren: Core, …)		B
14	Memory-/Core-Affinität		B
15	Accounting pro Nutzer, Nutzergruppe, ...		A
16	Redundanzmethode des Batchsystems		B
17	Für jeden Jobtyp (sequenziell, MPI, OpenMP, OpenMP+MPI) ist bei Installation ein Beispielskript		A

#	Forderung	Angebot	Kat.
	für einen einfachen Testjob mit einstellbarer Joblaufzeit zur Verfügung zu stellen		

5. Software-Support: Parallelisierung, Compiler, Bibliotheken

#	Forderung	Angebot	Kat.
1	Unterstützung für mindestens OpenMPI und MVAPICH2/Intel MPI		A
2	MPI optimiert zu den angebotenen Compilern, Bibliotheken und zum Infiniband		B
3	**Parallele Compiler-Unterstützung**		A
4	Unterstützung für GNU-Compiler		A
5	Intel Cluster Studio XE for Linux – Floating Academic 2-Seat License für mindestens 3 Jahre, Product with Maintenance and Support (2 Nutzer gleichzeitig auf Loginserver)		A
6	C/C++ (C/C++11)		A
7	Fortran 2003		A
8	Java 1.7		A
9	**Bibliotheken** (bevorzugt aus Intel-MKL oder Intel Compiler Suite)		A
10	BLAS		A
11	LAPACK		A
12	ScaLAPACK		A
13	Evtl. mit zur Verfügung gestellte Entwicklungstools für MPI und Shared Memory		I

6. Management des Gesamtsystems

#	Forderung	Angebot	Kat.
1	**Managementstation für gesamten Cluster**		A
2	(Angabe von CPU, Memory, Disk mit HW-Mirror, OS)		B
3	Mindestens 1 GbE-Management-Netzwerk		A
4	Typ, Switches, Ports des Management-Netzwerks		I
5	**Cluster-Management-Software** (Erfahrung mit XCAT, Salt Stack/Cobbler; herstellerunabhängige, bei Upgrades kostengünstige Managementsoftware bevorzugt)		A
6	*Single point of command* für alle Compute-, Login-Nodes, Fileserver, Disk-Array, IB (Bedienung: lokal über Managementstation und remote über RZ-LAN) inklusive folgender Funktionen:		A
7	Ein- und Ausschalten, Hoch- und Herunterfahren aller Systeme		A
8	Hardware-Monitoring und -Alerting		A
9	System-Monitoring und -Alerting (Erfahrung mit Graphite/Collectd/Grafana, Munin)		A
10	Schutz vor thermischer Überlast: automatisches Herunterfahren des gesamten Clusters bei kritischen (möglichst konfigurierbaren) Temperaturschwellenwerten		B
11	Hoch- und Herunterfahren des Clusters in richtiger und voll funktionstüchtiger Reihenfolge via Management-Software oder auf Wunsch automatisch konfigurierbar (z.B. nach Wiedererreichen der Normaltemperatur)		B
12	Installation der kompletten Systeme		B
13	Funktionsweise der Patch- und Releasepflege des gesamten Clusters		b
14	Komplette Technologie für diskless Compute-Nodes: Installation, Upgrades, Bootdauer einzelner und aller Nodes (Zeitangabe); Recovery einzelner und aller Nodes		A
15	Möglichkeit der Netzwerkkonfiguration über Cluster-		I

#	Forderung	Angebot	Kat.
	Management-Software (bitte verwendetes Tool nennen, falls es von CMS verschieden ist)		
16	Sicherung und Wiederherstellung aller Systeme des Clusters: • Fileserver, Login-Nodes, Cluster-Netzwerke • erzeugte Systemabbilder sollten zusätzlich via TSM-Backup in vorhandene ATL gesichert werden • Managementstation selbst • Filesystem und RAID		B
17	Nutzerauthentifizierung via LDAP		A
18	Möglichkeit der Anbindung an Active Directory		I
19	Beibehaltung des *single point of command* bei perspektivischem Ausbau des Clusters mit ca. 50 % Computenodes und 100 % Plattenplatz (ohne Komponenten auszutauschen)		B

7. Infrastruktur des Gesamtsystems

#	Forderung	Angebot	Kat.
1	Alle Racks und Kabel betriebsbereit montiert (Kabelführung über Kopf; sonstige lokale Gegebenheiten sind zu berücksichtigen) Anzahl Racks, Rackgröße		A
2	Elektrische Parameter: Genaue Anschlusswerte		I
3	Steckertyp und -anzahl pro Rack, die vom Kunden zur Verfügung zu stellen sind (Angabe notwendiger/nicht notwendiger Redundanzen)		I
4	**Maximale Energieaufnahme in KW des gesamten Clusters unter Volllast, inkl. Netzteilverluste** (alle Komponenten des Clusters sind im Zustand „Power On“ zu messen, wie in Abschnitt 9, Zeile 2)		B
5	Einhaltung der Obergrenze von 200 KW maximaler Energieaufnahme		A
6	Offenlegung der Ermittlungsmethode und Kalkulation, falls nicht auf einem identischen Cluster wie angeboten gemessen wurde, inklusive Angabe eventuell notwendiger Toleranzen aus Sicht des Anbieters		I
7	TFLOPS pro KW		I

Angaben des Kunden bei vorhandenen oder bauseits zu stellenden Klimaracks (Die Platzierung der vom Kunden zur Verfügung gestellten Kühlracks sind jeweils nach 2 Systemracks des Clusters einzuplanen)		
8	Typ der Kühlracks	Stulz CyberRow CRS 560 CW
9	Kühlleistung	39,2 kW/Rack (total), 26 kW/Rack nutzbar auf Grund von Redundanz und Wassertemperatur
10	Abmaße der Kühlracks (HxBxT, in mm)	1950 x 600 x 1175
11	Maximal verfügbare Stellfläche insgesamt mit und ohne Kühlracks	Platz und Kühlung für max. 10 Cluster-Racks geplant; siehe Lageplan auf itz.uni-

Angaben des Kunden bei vorhandenen oder bauseits zu stellenden Klimaracks (Die Platzierung der vom Kunden zur Verfügung gestellten Kühlracks sind jeweils nach 2 Systemracks des Clusters einzuplanen)		
		halle.de/hpc/ausschreibung_2015
12	Maximale Belastbarkeit des Doppelbodens	500 kg/m² („Schwerlastboden“)

Platz

8. Gewährleistung, Service, Support, Hotline, Installation, Inbetriebnahme des kompletten Clusters (Abschnitte 1–7) und deren einwandfreies Zusammenspiel aus einer Hand

#	Forderung	Angebot	Kat.
1	5 Jahre Gewährleistung auf alle Komponenten: Hardware, Firmware, Systemumgebung; aber 3 Jahre Gewährleistung auf Compute-Nodes. Ersatzteile und Technikereinsatz vor Ort am nächsten Arbeitstag für Abschnitte 1–7		A
2	Garantierte Verfügbarkeit passfähiger Ersatzteile mit mindestens identischer Funktion im Bedarfsfall für mindestens 5 Jahre nach erfolgter Übergabe des Clusters		A
3	Örtliche Entfernung des zuständigen Servicepersonals (zur Abschätzung der Reaktionszeit bei Vor-Ort-Einsatz im Dringlichkeitsfall)		B
4	10 Tage Consulting flexibel nutzbar nach Übergabe des Clusters für folgende Bedingungen innerhalb von 5 Jahren: 5 Jahre Upgrades und Support inkl. How-To-Do Hotline, inkl. Analyse der Passfähigkeit aller Komponenten: Firmware, Systemumgebung (OS, paralleles FS, Batchsystem, Cluster-Management, IB, Disk-Array (siehe Abschnitte 1–7)		B
5	Mehrkosten für 20 Tage Consulting		b

9. Benchmarks

Alle Benchmarks mit Anmerkungen finden Sie zum Download unter itz.uni-halle.de/hpc/ausschreibung_2015.

#	Forderung	Angebot	Kat.
1	**Basisbenchmarks zur Bewertung und Überprüfung grundlegender Funktionen, der Performance und des maximalen Stromverbrauchs** Offenlegung der Ermittlungsmethode und Kalkulation, falls nicht auf einem identischen Cluster wie angeboten gemessen wurde, inkl. Angabe eventuell notwendiger Toleranzen aus Sicht des Anbieters.		B
2	LINPACK mit Turbo on (in TFLOPS und in % der Peak-Performance) als Funktionstest für das gesamte Cluster (alle Komponenten des Clusters sind im Zustand power on); die verwendeten Bibliotheken, Compiler und Compileroptionen sind anzugeben: • 50 % der Nodes, alle Cores mit IB 1:2 und 50 % IB 1:8 (2 Messungen gleichzeitig) • alle Cores an einem Switch (als eine „Insel“) und alle „Inseln“ gleichzeitig (hierbei Überprüfung der maximalen Energieaufnahme inkl. Netzteilverluste wie vom Anbieter in Abschnitt 7, Zeile 5 genannt)		B
3	SpecFP-Rate pro Node (Angabe aus Datenblatt mit Quellenangabe zulässig)		B
4	STREAM Triad (in GB/s) pro Nodetyp		B
5	HPCC Benchmark-Suite		B
6	IOR: maximal erreichbare IO-Bandbreite für das parallele FS (in GB/s)		B
7	IOR: maximal erreichbare IO-Bandbreite für ein Compute-Node (in GB/s)		B
8	Mdtest: Metadaten-Performance des Dateisystems		B
9	**Kundenspezifische Benchmarks**		B
10	Benchmark Physik/Chemie: GROMACS		B

#	Forderung	Angebot	Kat.
11	Benchmark Pharmazie: AMBER, Weka		B
12	Benchmark Life Science: Samtools, BWA, BLAST		B
13	Benchmark Life Science: Off-target Prediction		B
14	Benchmark Life Science: Sequence Trimming		B
15	Benchmark Life Science: HMM		B

Vor der Übergabe des Systems sind die Benchmark-Werte des Angebots mit dem Kunden vor Ort nachzuvollziehen und zu dokumentieren.

10. Dokumentation

#	Forderung	Angebot	Kat.
1	**Komplette Dokumentation mit folgenden Inhalten bei der Übergabe:**		A
2	Installation und Konfiguration aller HW- und SW-Komponenten (außer Anwendungen)		A
3	alle Recovery-Szenarien aller Komponenten des gesamten Clusters		A
4	Tests und Benchmarks aus Abschnitt 9 „Benchmarks“ müssen durch den Kunden nachvollziehbar sein		A
5	Batch-Beispielskripte (siehe Abschnitt 4, Zeile 17)		A
6	konfigurierbare Filter sämtlicher Logfiles mit Beispiel nach Kundenvorgaben		b

11. Anlagen

Folgende Anlagen sind unbedingt beizulegen:

1. Benchmark-Ergebnisse
2. Stromverbrauch und Klimabedarf (pro Komponente und insgesamt), möglichst tabellarisch

Angebotsgültigkeit

Datum Name Unterzeichner Unterschrift Stempel