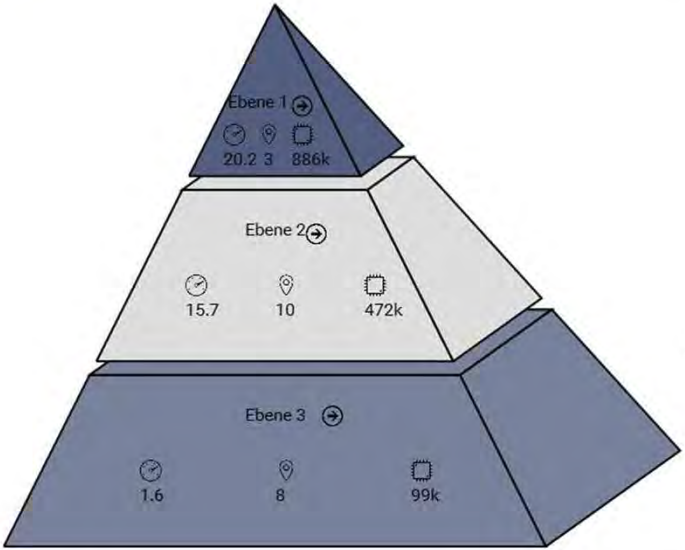
Rechnerbeschaffung

1. HPC-Versorgung in Deutschland
2. Phasenmodell Beschaffung
3. Antragstellung
4. Markerkundung und Ausschreibung
5. Vertragsverhandlungen
6. Rechnerraumumbau und Installation
7. Produktionsbetrieb

HPC-Versorgung in Deutschland



DKRZ: 3,6 PFLOPS, 100.000 Kerne

Januar 2017

Versorgungspyramide HPC

- „Höchstleistungs"rechner (Ebene 1)
  - Versorgung für Europa, Bund und Land
  - Gauss Center for Supercomputing
  - LRZ (Garching), HLRS (Stuttgart), JSC (Jülich)
- „Hochleistungs"rechner (Ebene 2)
  - Versorgung für Bund und Land
  - Z.B. Dresden, Aachen, Darmstadt, Hamburg (DKRZ)
- „Hochleistungs"rechner (Ebene 3)
  - Versorgung Land
  - Z.B. Erlangen, Kaiserslautern

Geographische Verteilung

- Nicht alle Bundesländer vertreten
- Nordbundesländer im HLRN zusammengefasst
- Schwerpunkte in den industriereichen Bundesländern

Finanzmittel

Ergebnis politischer Entscheidungen

- Ebene 1
  - Aktuell ca. 130 M€ pro Beschaffungszyklus und System
  - Beinhaltet auch Betriebskosten und Rechnergebäude
- Ebene 2
  - Typischerweise 15+ M€ pro Beschaffungszyklus und System
  - Ausnahme DKRZ mit 40+ M€ pro System
    - Andere Ausnahmen für Zentren mit anderen Finanziers
- Ebene 3
  - Typischerweise einige M€ pro Beschaffungszyklus und System

Künftige neues Finanzierungskonzept im Rahmen des Nationalen Hoch- und Höchstleistungsrechnens (NHR)

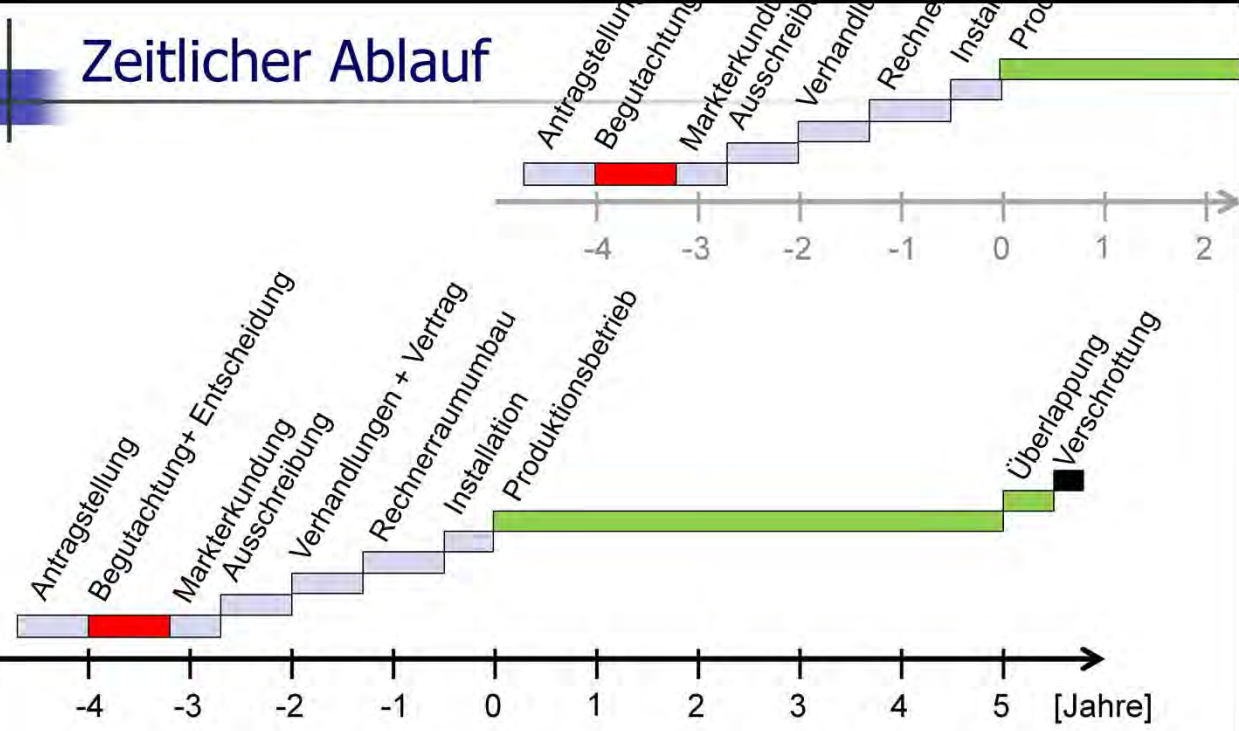
Phasenmodell Beschaffung

Regelmäßige Beschaffung z.B. alle 5 Jahre

- Antragstellung
- Markterkundung
- Ausschreibung
- Verhandlungen
- Kaufentscheidung
- Rechnerraumumbau
- Installation
- Produktionsbetrieb

- Dauerhaft:
  Politische Aktivitäten

Zeitlicher Ablauf



Antragstellung

- Absehbar in 4 Jahren
  - Alter Rechner überlastet (DKRZ: ca. 4fach überbucht)
  - Wissenschaft nicht mehr optimal unterstützt
- Am Anfang steht das Geld
  - Festlegung eines Rahmens liegt meistens vor
  - Üblicherweise: kontinuierliche Weiterentwicklung der Zentren, ihrer Systeme und Dienste
- Struktur eines Antrags (50-150 Seiten)
  - Darstellung neuer wissenschaftlicher Ziele (=Bedarf)
  - Darstellung neuer technischer Möglichkeiten im HPC (=Möglichkeiten der Bedarfsdeckung)
  - Abschätzung von verfügbarer Hardware, ihrer Beschaffungs- und Betriebskosten (=Umsetzungsplan)
  - Beschaffungs- und Betriebskonzept (=Details)

Neue wissenschaftliche Ziele

DRKZ und Klimaforschung: relativ homogenes Profil

- Neue Methoden der Wissenschaftler
  - Höhere räumliche und zeitliche Auflösung der Modelle
  - Mehr Prozesse (Wolken, Chemie und anderes)
  - Mehr Ensemble-Mitglieder
    Ensemble-Berechnungen: Verrechnungen statistischer Schwankungen der Ergebnisse bei modifizierten Eingaben

Allgemeine Rechenzentren

- Verschiedene Wissenschaftsbereiche mit unterschiedlichen Programmcode-Strukturen und Abläufen in der computerbasierten Modellierung und Datenauswertung
  - Teilweise nur kleinere Datenmengen benötigt
  - Manchmal gut auf Beschleunigerhardware implementierbar/portierbar
  - ...

Neue wissenschaftliche Ziele...

Darstellung im Antrag

- Künftige Bedarfe an
  - Rechenzeit
  - Speicherplatz
  - Spezial-Hardware (Beschleunigung, Visualisierung, Nachverarbeitung...)
- Erwünschter Ausbau am DKRZ: Faktor 10...100

Problem

- Wir befinden uns **4 Jahre** vor Inbetriebnahme des Systems
- Wissenschaftsentwicklung schwer vorhersagbar

Neue technische Möglichkeiten

Entwicklung in der Hardware

- Prozessoren
  - Prozessorfamilien von Intel; sonst noch Firmen?
- Speichersysteme
  - Insbesondere Dateisysteme: Lustre oder GPFS?
  - SSD-Burst-Buffer? Non volatile memory?
- Vernetzungen
  - Infiniband; wenige Überraschungen
- Spezialhardware
  - Beschleuniger: GPGPU, FPGA, Xeon Phi ...
  - Visualisierung
  - Bandspeicherung (HSM)

Problem

- Wir befinden uns **4 Jahre** vor Inbetriebnahme des Systems
- Technische Entwicklung schwer vorhersagbar

Abschätzung der verfügb. HW und Kosten

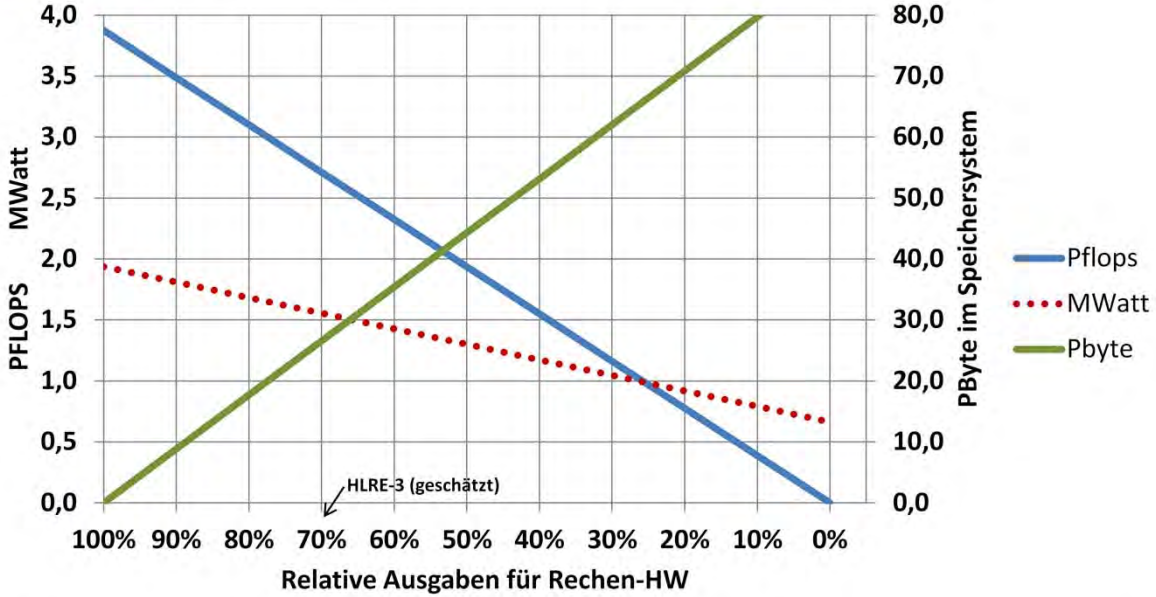
Beschaffungs- und Betriebskosten

- Aus getrennten Budgets: Beschaffung durch Beantragung, Betrieb aus Jahreshaushalt
- Muss am Ende zusammenpassen
- Problem: wenn ich alles Geld für HW ausgebe, kann ich dann den Strom des Systems bezahlen?

Speziell am DKRZ

- Aufteilung der Ausgaben für Rechnen und Speichern

Abschätzung der verfügb. HW und Kosten...



Abschätzung der verfügb. HW und Kosten...

Abschätzung HLRE-3 im Antrag 2011

Rechnersystem: 31M€ - HSM: 5M€ - Umbauten: 5M€

Characteristic	HLRE-2	HLRE-3	Factor
a) Peak performance	158 TFLOPS	~ 3,000 TFLOPS	~ 20
b) No. of processor cores	8,300	~ 120,000	~ 14
c) Mean memory per core	3 GB (cores with 4GB and 2GB)	~ 3 GB (cores with 4GB and 2GB)	~ 1
d) Main memory	20 TB	~ 360 TB	~ 18
e) Storage on disk	6 PB	~ 120 PB	~ 20
f) Memory-to-disk	30 GB/s	~ 600 GB/s	~ 20
g) Full memory dump to disk	< 1 hour	< 1 hour	~ 1
h) Storage on tape	65 PB	~ 650 PB	~ 10
i) Disk-to-tape	3 GB/s	~ 30 GB/s	~ 10
j) Annual data production	10 PB/year	~ 100 PB/year	~ 10
k) Overall power consumption	2 MW	2 MW	1

Beschaffungs- und Betriebskonzept

Mehrphasige Installation

- Installation in zwei Phasen mit einem Jahr Abstand
- Ziel
  - Im ersten Jahr ist der Rechner noch nicht ausgelastet
    - Kleineres System genügt; wir sparen Strom
  - Wir sparen Geld auf für bessere Technik
    - HLRE-3: Haswell- und Broadwell-Prozessoren
  - Nachteil: System ist in den Komponenten heterogen
    - Erschwerte Verwaltung, Jobs eher nicht auf beiden Teilen zugleich
- Alternativen
  - Z.B. mehrere Systeme, die im Wechsel oder in Ergänzung hochgezogen werden (z.B. in Jülich)
  - Nachteil: wiederholte vollständige Beschaffungen notwendig

Markterkundung

Zeitpunkt: 3 Jahre vor Inbetriebnahme

Ziel

- Frühzeitige Kontaktaufnahme mit potentiellen Anbietern
- Übersicht über Entwicklungslinien bei
  - Prozessoren
  - Speichersystemen
  - Anderen HW- und SW-Systemen
- Grobe erste Ideen von PFLOPS/M€ und PByte/M€
- Erste Abschätzungen von Stromverbrauchen
- Kommunikation unserer Zielvorstellungen an Hersteller

Wichtig: Ausschreibung muss damit umsetzbar sein

Ausschreibung

- Zeitpunkt: 2 Jahre vor Inbetriebnahme
- Europaweite Ausschreibung nach den Regularien aus dem öffentlichen Bereich
  - Strenge rechtliche Abwicklungsvorgaben zur Erzielung von Chancengleichheit, Korruptionsfreiheit usw.
  - Vermeide IT-Elbphilharmonie ☺
- Üblich bei anderen Produkten
  - Bedarf definieren – Ausschreibung gewinnt der Bieter mit dem wirtschaftlichsten Angebot (nicht notwendigerweise das billigste)
- Üblich bei HPC-Ausschreibungen
  - Geldsumme festlegen – Ausschreibung gewinnt der Bieter mit der am besten bewerteten Leistung (oft: meiste Hardware)
- Bewertungsschema mit Ausschreibung festgeschrieben

Ausschreibungsdokument

Vergabeunterlagen (RFP – Request for Proposals)

- Festlegung der Wertung der Angebote
- Systempreis und Preise für Erweiterungen
- Leistungsanforderungen Rechnen (Phase 1 und 2)
- Leistungsanforderungen Speichern (Phase 1 und 2)
- Unterstützende HW und SW
- Elektrische Leistungsaufnahme
- Integration in bestehende Infrastruktur
- Benchmarks

Umfang: ca. 50 Seiten am DKRZ

Ausschreibungsdokument...

Festlegung der Wertung der Angebote

- Als Excel-Tabelle mit etwa folgenden Gewichtungen
  - Preis (1/4 der Punkte)
  - Preise für Erweiterungen (TFLOPS, PByte)
  - Rechenleistung (ca. 1/3 der Punkte)
  - Speichersystem (ca. die Hälfte der Rechenleistung)
  - Punkte für Software und andere (weiche) Faktoren
- Die Ausschreibung gewinnt der Bieter mit der höchsten Punktezahl
- Muss alles rechtssicher dokumentiert werden
- Klagen bei Fehlern sind wahrscheinlich!

Ausschreibungsdokument...

Leistungsanforderungen Rechnen (Phase 1 und 2)

- Integrale Leistung über beide Phasen angefordert
- Leistungsbewertung aufgrund von Benchmarks

Ausschreibungsdokument...

Leistungsanforderungen Speichern (Phase 1 und 2)

- Speicherkapazitäten z.B. vorgegeben
  - Für beide Phasen getrennt
- Anzahl zu speichernder Dateien vorgegeben
- Forderung nach qualitativem Dateisystem
  - Z.B. Zeit zum Neustart nach Absturz
- Problem: es gibt nur Lustre und GPFS
  - IBM liefert GPFS, alle anderen Lustre

Ausschreibungsdokument...

Unterstützende Hardware und Software

- Zusätzliche Rechnerknoten mit Grafikkarten für Visualisierungen
  - Direkt angebunden an das Speichersystem
- Testsystem
  - Für Tests neuer Software und Firmware
- Software-Komponenten
  - Linux
  - Batch-Scheduler mit Vorgabe bzgl. Steuermöglichkeiten
  - Backup-Software
  - Compiler Fortran/C/C++, Bibliotheken, MPI, OpenMP
  - Werkzeuge zur Fehlersuche und Leistungsanalyse

Ausschreibungsdokument...

Elektrische Leistungsaufnahme

- Vorgabe eines maximalen akzeptierten Verbrauchs
  - Grund: Finanzierung von Strom und steigenden Energiesteuern ist kritisch
- Ermittelt durch Mix von realistischen Benchmark-Programmen
  - LINPACK nur geeignet, um Maximalverbrauch zu testen
  - Hier
    - Maschine mit Anwendungsbenchmarks vollpacken
      Dann Leistungsaufnahme messen

Ausschreibungsdokument...

Integration in bestehende Infrastruktur

- Vorgabe der Stellflächen
- Vorgabe der Bodenbelastungen
- Vorgabe der Stromschienen und Stromverteiler
- Vorgabe der Kühlsysteme

Weitere Kleinigkeiten

- Deckenhöhen
- Säulenabstände
- Türweiten und Belastbarkeit der Aufzüge

Ausschreibungsdokument...

Leistungsanforderungen aus Kundensicht

- Anzahl geeigneter Bieter: ca. 3-6 Hersteller
- Was hindert Bieter an einer Teilnahme?
  - Z.B. zu strenge Vorgaben für Dateisystem
  - Z.B. zu enges Strombudget
- Folge: Ausschreibung bleibt ohne Angebote
  - Ist an anderer Stelle bereits geschehen
  - Neuausschreibung erforderlich
    - Zeitverlust, Reputationsverlust
    - Probleme mit Finanzierung von altem und neuem Rechner

Ausschreibungsdokument...

Leistungsanforderungen aus Bietersicht

- Bieter bietet viel Leistung
  - Höhere Gewinnchancen im Wettbewerb
  - Muss dann aber auch die nötige Hardware liefern, wenn er gewinnt
- Bieter ist vorsichtig mit Leistungsprognosen
  - Verringerte Gewinnchancen
  - Im Gewinnfall aber auch realistische Hardware-Lieferung

IT-Branche allgemein: schwieriges Geschäft

- Beteiligung an Ausschreibung teuer für Bieter

Ausschreibungsdokument...

Benchmarks

- Benchmarking ist eine Kunst
- Es gibt unzählige Vorgehensvarianten
- Vielleicht einfachste: LINPACK-Benchmark verwenden
- Unser Ansatz
  - Rechenleistung
    Anwendungsbenchmarks
    Mix relevanter Modelle der Kunden mit MPI und OpenMP
    Methode: Erhöhung des Jobdurchsatzes
  - Speichersystemleistung
    Synthetische Benchmarks
    Methode: Vorgabe von Leistungsdaten

Ausschreibungsdokument... Benchmarking

Anwendungsbenchmarks – Vorgehensmodell (1)

- Wir bestimmen eine Referenzlaufzeit für einen Modellcode, z.B. 10 Minuten (dazu benötigen wir n Kerne)
- Wir ermitteln auf unserer alten Maschine, wieviele Jobs wir pro Sekunde auf der vollen Maschine durchbekommen
- Der Bieter ermittelt die kleinste Anzahl von Kernen, mit denen er die Referenzzeit unterschreitet
- Für die von ihm gebotene Anzahl Kerne bestimmt er den Durchsatz für seine volle Maschine
- Der Quotient der beiden Durchsätze ist die Durchsatzsteigerung für diesen Benchmark

Ausschreibungsdokument... Benchmarking

Anwendungsbenchmarks – Vorgehensmodell (2)

- Nicht ein Benchmarkcode sondern ein halbes Dutzend
- Jeweils evaluiert in zwei Varianten
  - Unoptimiert (nur Compilereinstellungen)
    Zeigt uns, was das System und der Compiler können
  - Optimiert
    Zeigt uns, was das Team des Bieters leisten kann
- Macht ein Dutzend Varianten
  - Unterschiedliche Gewichtung pro Benchmark (optimierte geringer)
- Getrennt angegeben für Phase 1 und Phase1+Phase2
- Gewinner ist der Bieter mit dem höchsten Wert

Ausschreibungsdokument... Benchmarking

E/A-Benchmarks

- Single stream Posix-Transferrate von/zu Rechnerknoten
- Aggregierte Posix-Streaming-Transferrate
- Leistungen für HDF5 und NetCDF
- Parallele E/A mit MPI auf eine Datei
- Metadaten-Leistung

Hier jeweils Vorgaben an den Bieter, die er einhalten muss

Vertragsverhandlungen

Zeitpunkt: ca. 1,5-2 Jahre vor Inbetriebnahme

- Bieter liefern Angebot
  - Dokument mit 200+ Seiten
    - Technische Spezifikation
    - Ergebnisse des Benchmarking
    - Konditionen für Lieferung, Inbetriebnahme, Wartung usw.
- Typischerweise stellen beide Seiten Problembereiche im Ausschreibungsdokument fest
  - Weitere Detaillierung der Ausschreibung
  - Kommunikation an alle
  - Neue Runde mit Angeboten
- Nach mehreren Runden konvergiert das Verfahren

Ergebnisse des Benchmarking ☺

Probleme

- Anwendungsbenchmarks: Zielprozessor existiert nicht
  - Alle Angaben sind Hochrechnungen
  - Erstellt auf einem existierenden Prozessor (Vorgängermodell)
  - Hintergrundinformation des Prozessorherstellers über Leistungszuwachse der kommenden Generation am Bieter
  - Hochrechnungen und Simulationen beim Bieter
  - Datum der Markeinführung nicht genau bestimmbar
  - Varianten des Prozessor bei Markteinführung im Voraus nicht final bekannt
  - Preise auch nicht final bekannt
- Somit: Bieter spielt Benchmark-Poker
- Verfehlen der Leistungszusagen
  - Erfolgreicher Bieter muss soviel HW nachliefern, bis Leistungszusage erfüllt

Ergebnisse des Benchmarking... ☺

Probleme

- E/A-Benchmarks: System der ausgeschriebenen Größe existiert noch gar nicht
  - Leistungsangaben sind Hochrechnungen
  - Insbesondere bei Dateisystemen unklar, ob sie die geforderte Größe und Skalierbarkeit mit allen Qualitätsforderungen erfüllen können
- Somit: Bieter spielt Benchmark-Poker
- Verfehlen der Leistungszusagen
  - Erfolgreicher Bieter muss soviel HW nachliefern, bis Leistungszusage erfüllt

Rechnerraumumbau und Installation

Rechnerraumumbau und Installation...

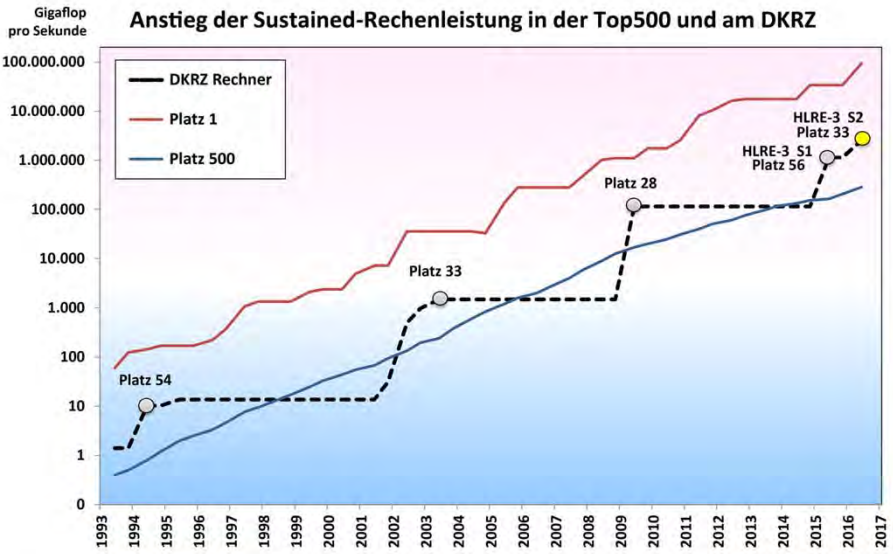
Umbauten am Rechnerraum (HLRE-3)

- Stromversorgung
  - Batteriepufferung (höhere Verfügbarkeit)
  - Weiterer Mittelspannungstransformator (höhere Verfügbarkeit)
  - Umbau der Stromschienen (andere Aufstellung im Raum)
- Kühlung
  - Rechnersysteme mit Hochtemperaturflüssigkeitskühlung
    - Ermöglicht ganzjährige freie Kühlung über das Dach ohne weitere Kühlaggregate
- Bandarchiv mit Sauerstoffreduktionsanlage
  - Reduktion von 20,5% (normal) auf 17% und 15%
  - Entstehung von Bränden weitestgehend verhindert

Batteriepufferung

Produktionsbetrieb

DKRZ ist wieder im Rennen



Danach geht alles von vorne los...

Rechnerbeschaffung
Zusammenfassung

- Zeitdauer der Beschaffung: >4 Jahre
- Antragstellung sehr früh mit erster Abschätzung zu Wissenschaft und HPC-Technik
- DKRZ: Aufteilung der Finanzmittel auf Rechnen und Speichern wichtig und schwierig
- Beschaffung mit zwei Installationsphasen
- Ausschreibung ist aufwendig
- Bieterverfahren muss rechtskonform ablaufen
- Problem für Kunde: Definition geeigneter Benchmarks
- Problem für Bieter: Benchmark-Hochrechnungen bei nicht-existierender Ziel-Hardware

Rechnerbeschaffung
Die wichtigsten Fragen

- In wievielen Phasen soll die Installation ablaufen?
- Wie könnte ein Benchmarking für eine zu beschaffende Maschine aussehen?
- Was muss der Kunde beachten?
- Welche Probleme stellen sich für den Bieter?