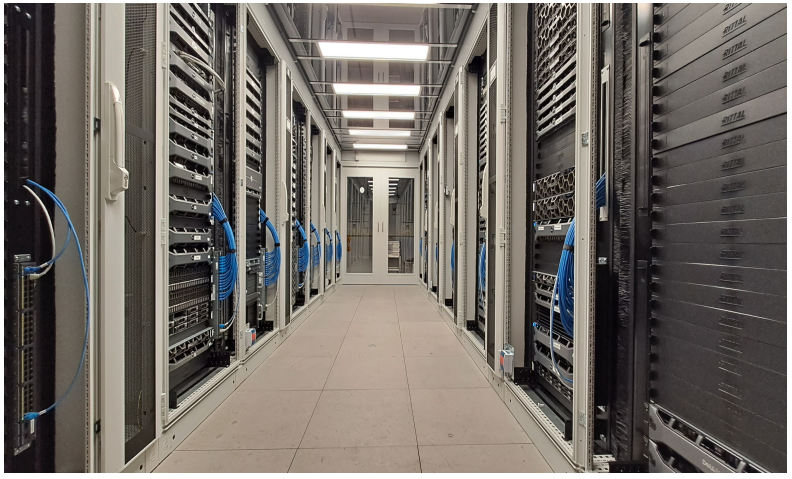
Das Zeuthener Rechenzentrum



Kai Leffhalm

Campus Seminar

11. Juli 2024

Desy Computing Center: Herkunft

Geschichte

1963 Erste Lochkarten Maschine zum Speichern der Spuren Messdaten der Blasenkammer

1969 Schnellere Maschine, einlesen und Speichern von Lochkarten

1970 Mehr Bit, mehr Speicher, erste Programmiersprache, Datenaustausch mit anderen Instituten

1981 Neue Programmiersprachen, Softwareentwicklung für die Datenauslese

1987 Vom INFN entwickelte Hardware Ape1 für Gittereichtheorie

1989 Simulationen, Datenaustausch per Magnetband

1994 IBM Rechner

1994 2. Version: APE100

1996 Tape Roboter

2000 3. Version mit Entwicklungsbeiträgen von Desy: ApeMille

2005 4. Version: Gemeinsame Entwicklung von INFN und Desy: ApeNext

2008 Austausch Tape Roboter

200X Standard Hardware Server von Dell und HP

2009 Zweiter Rechnersaal

2011 ca. Einrichtung erste Kaltgangeinhausung

Unser Rechenzentrum - Die Basis für High Throughput Computing

Netzwerkzugang

||

Gastzugang

- 4x10Gb/s über dedicated Fibre von Desy in Hamburg
- Firewall schützt internes Netzwerk
- Extra OpenInfra Leitung für Gästenetze, Firmennetze
  - extra Firewall

Unser Rechenzentrum - Die Basis für High Throughput Computing

⇒

Netzwerkzugang

- Zugang per ssh, vpn mit 2. Faktor
- Zugang aus den Büros und aus dem Desy-WLAN ohne 2. Faktor
- Per VPN direkter Zugang zu Windows Desktop mit RDP

Unser Rechenzentrum - Die Basis für High Throughput Computing

Netzwerkzugang

⇒

WGS

- Work group server als interaktive Rechenknoten
- Gruppenweise eigene WGS
- Submit Hosts für computing Farm

Unser Rechenzentrum - Die Basis für High Throughput Computing

Netzwerkzugang

⇒

WGS

⇒

lustre

dCache

afs

- Mass storage auf dCache: 70 Knoten mit 12 PB
- lustre verteilter Storage auf 27 Knoten mit 3,6 PB
- Multifunktionaler Storage afs mit 270 TB, sehr alt
- Bald auch Ceph mit S3 und weiteren Möglichkeiten

Unser Rechenzentrum - Die Basis für High Throughput Computing

Netzwerkzugang

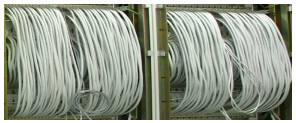
⇒

WGS

⇒

Grid

Farm



Parallel

GPU

- Grid Computing: 4200 Kerne für weltweiten Zugriff
- Farm Computing (HTC): 3360 Kerne für lokale Nutzer
- Parallel Computing (HPC): 2048 Kerne für lokale Nutzer
- GPUs: 120 GPUs integriert in die Farm für lokale Nutzer

Unser Rechenzentrum - Die Basis für High Throughput Computing

Netzwerkzugang

⇒

WGS

⇒

Grid

Farm,Parallel,GPU

⇒

dCache

dCache, lustre, afs

- Zugriff über Grid auf dCache Instanz für Grid
- Zugriff über WGS und Farm auf die lokalen Storage Instanzen

Unser Rechenzentrum - Die Basis für High Throughput Computing

Netzwerkzugang

⇒

WGS

⇒

Grid

Farm,Parallel,GPU

⇒

dCache

dCache, lustre, afs

⇒

Tape

- Tape Anbindung nur über lokale dCache Instanz
- Tape Roboter sind in der Anschaffung teurer,
  - lohnen sich erst ab einer gewissen Menge an Daten
- Stromverbrauch ist aber konstant bei zur Zeit 1,1 kW, demnächst bei 0,6 kW

Die unterschiedlichen Storagesysteme

dCache

- Am Desy mitentwickeltes Storage System
- Als Cache vor einem Tape System gestartet
- Kann mittlerweile viele verschiedene Protokolle sprechen
- Jeder Poolknoten hat direkte Verbindung zum Klienten
- Migration von Pool zu Pool durch Admins

lustre

- Storage System Entwicklung mit Schwerpunkt auf HPC
- Wird als Block installiert, Migration durch Nutzer
- Ist offiziell ein „scratch“ System

afs

- Ältestes File System im Haus
- Sehr gute Technik, Weiterentwicklung aber unklar
- Homedirs auf Linux, Gruppenverzeichnisse, Webverzeichnis, Softwareinstallationen

Über die Basis hinaus

- Mit der großen Anzahl an Servern ergaben sich viele Möglichkeiten
- Nutzer wollten mehr als „nur“ Computing-Power

Über die Basis hinaus

- Mit der großen Anzahl an Servern ergaben sich viele Möglichkeiten
- Nutzer wollten mehr als „nur“ Computing-Power

Virtuelle Maschinen

⇒

WGS

⇒

Datenbanken

Spezielle Hardware

Über die Basis hinaus

- Mit der großen Anzahl an Servern ergaben sich viele Möglichkeiten
- Nutzer wollten mehr als „nur“ Computing-Power

Virtuelle Maschinen

Entwicklungstools

⇒

WGS

⇒

Datenbanken

⇒

Weltweiter Zugang

Spezielle Hardware

Root Zugang

Über die Basis hinaus

- Mit der großen Anzahl an Servern ergaben sich viele Möglichkeiten
- Nutzer wollten mehr als „nur“ Computing-Power

Virtuelle Maschinen

Entwicklungstools

Kubernetes

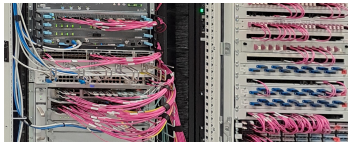
⇒

WGS

⇒

Datenbanken

⇒



Weltweiter Zugang

⇒

DMZ

Spezielle Hardware

Root Zugang

Abgeschottet

Entwicklungsumgebungen

GitLab CI

- Instanz ist in HH
- Globaler gitlab runner ist verfügbar
- Gitlab runner für spezielle Anwendungen möglich

Jenkins

- Harbor wir als container repository genutzt
- sonarqube ist verfügbar
- nexus für artifacts und als proxy verfügbar

Services direkt für Nutzer

Kubernetes
- Container Umgebung für spezielle Anforderungen
- . . . AMPEL
- . . . bald ILDG Meta Data Katalog
- . . . Redmine Projekt Management
- . . . CTAO Services

S3 object store
- Tests Ceph Storage System
- Mit extra Hardware auch per MinIO

InfluxDB
- Bienen Project
- MQTT als message queue verfügbar
- wird zum Auslesen von Sensor Daten genutzt

jupyter
- Für Tests verfügbar
- Zur Zeit nur ein Server
- . . . und nur intern erreichbar
- Unsere Storage Systeme sind verfügbar

Services direkt für Nutzer

Kubernetes
- Container Umgebung für spezielle Anforderungen
- . . . AMPEL
- . . . bald ILDG Meta Data Katalog
- . . . Redmine Projekt Management
- . . . CTAO Services

S3 object store
- Tests Ceph Storage System
- Mit extra Hardware auch per MinIO

InfluxDB
- Bienen Project
- MQTT als message queue verfügbar
- wird zum Auslesen von Sensor Daten genutzt

jupyter
- Für Tests verfügbar
- Zur Zeit nur ein Server
- . . . und nur intern erreichbar
- Unsere Storage Systeme sind verfügbar

Interessiert? Schreibt an uco-zeuthen@desy.de

Die Netzbereiche

DMZ - Demilitarisierte Zone

- Abgetrennte Netzwerkbereiche, die nicht vollen Zugriff auf internes Netz haben
  - . . . auch nicht untereinander
- Dort sind extern erreichbare Dienste
- Erfolgreicher Angriff nur auf wenige Systeme ausdehnbar

Abschottung

- Abschottung unsicherer Dienste ohne externen Zugang in extra Netzwerkbereichen
  - Hardware Entwicklung
  - Alte Geräte ohne Sicherheitsupdates
- Netzwerk kann Subnetze voneinander trennen oder nur eine Richtung des Zugangs erlauben
  - Wichtige zentrale Services sind nur von wenigen Systemen aus erreichbar
  - Unsichere Systeme werden vor dem Internet abgeschottet
  - Wichtige Systeme werden von den unsicheren geschützt

Notwendige Infrastrukturen

- Einfach an die Steckdose und los . . . funktioniert natürlich nicht
- Redundanz und Ausfallsicherheit muss hoch sein
  - Verfügbarkeit und Zuverlässigkeit wichtiges Element eines Rechenzentrums

USV, 600 kW

Notwendige Infrastrukturen

- Einfach an die Steckdose und los . . . funktioniert natürlich nicht
- Redundanz und Ausfallsicherheit muss hoch sein
  - Verfügbarkeit und Zuverlässigkeit wichtiges Element eines Rechenzentrums

USV, 600 kW

Kühlung, 500 kW

Freikühler, 1 MW

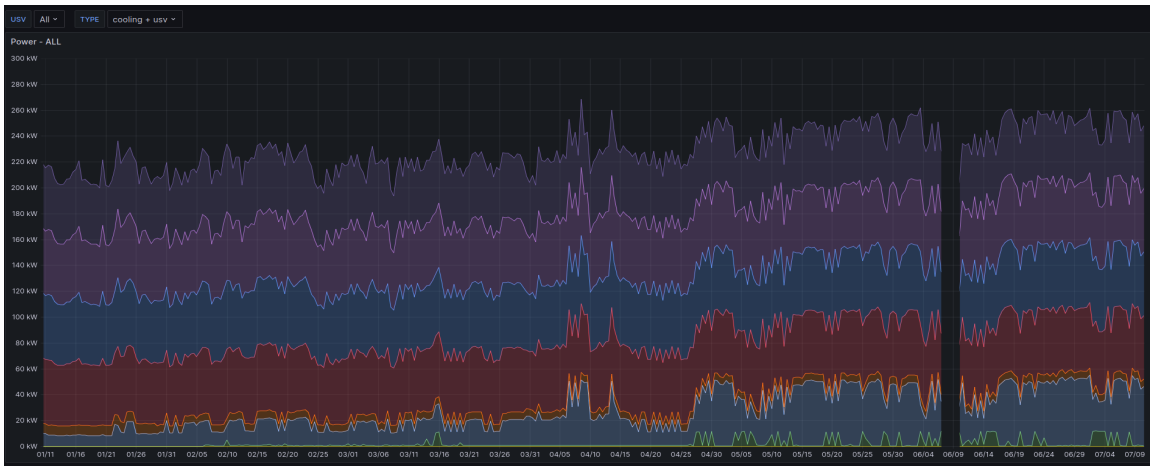
Notwendige Infrastrukturen

- Einfach an die Steckdose und los . . . funktioniert natürlich nicht
- Redundanz und Ausfallsicherheit muss hoch sein
  - Verfügbarkeit und Zuverlässigkeit wichtiges Element eines Rechenzentrums
- Kosten-Effizienz ist auch wichtig

USV, 600 kW

Kühlung, 500 kW

⇒



Stromverbrauch

Freikühler, 1 MW

- Kosten-Effizienz ist auch wichtig

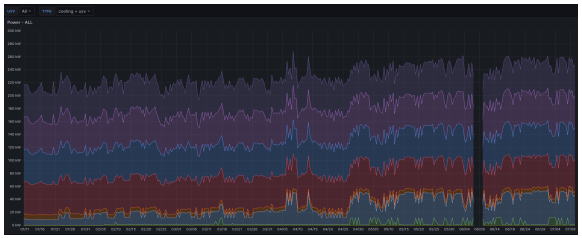
Notwendige Infrastrukturen

- Einfach an die Steckdose und los . . . funktioniert natürlich nicht
- Redundanz und Ausfallsicherheit muss hoch sein
  - Verfügbarkeit und Zuverlässigkeit wichtiges Element eines Rechenzentrums
- Kosten-Effizienz ist auch wichtig

USV, 600 kW

Kühlung, 500 kW

⇒



Stromverbrauch

⇒

Einhausung

Freikühler, 1 MW

- Kosten-Effizienz ist auch wichtig

Notwendige Technologien

- Datenspeicherung auf RAID Systemen
  - Backup der wissenschaftlichen Daten zu teuer
  - Nur Homedirs im Backup
- Enterprise Hardware
  - läuft 24h - 7 Tage die Woche - 5-10 Jahre durchgehend
  - gerade bei Festplatten wichtig, sonst droht Datenverlust und viel Austauscharbeit
  - Festplatten melden sich sogar selber, bevor sie kaputtgehen
- Stromnetzteile immer doppelt ausgelegt
- Stromanschluss an unterschiedliche Phasen
- Netzwerkanschlüsse doppelt ausgelegt
- Netzwerk an 2 Switche
- Alle Server und Netzwerkgeräte über internen Serviceprozessor remote konfigurierbar
  - Bios, Controller, . . .
  - Bei 700 Servern kann man nicht zu jedem Server hinlaufen
- Dienste sollten auch wenn möglich immer eine Ausfallabsicherung haben
  - Cold Standby
  - Hot Standby
  - HA Lösung

Aktuelle Bauprojekte

Alte Luftkühlung

⇒

Neue Kaltgangeinhausung

Aktuelle Bauprojekte

Alter Tape Roboter

⇒

Neuer Tape Roboter

Vielen Dank

Nutzer/Office Support

Simone, Patrick, Susann, Bruno, Mathias, Daniel, Lukas

Software Entwicklung

Hasmik, David, Bagrat, Davit, Stefan, Grygorii, Thomas, Dima

Vielen Dank
Fragen?

Netzwerk

Axel, Rico, Chris, Markus, Carsten

Ausbildung

Tim, Simon, Mithat, Sehyang

Scientific Computing

Stephan, Andreas, Götz, Ingo, Philipp, Holger, Vlad, Kai

Backup

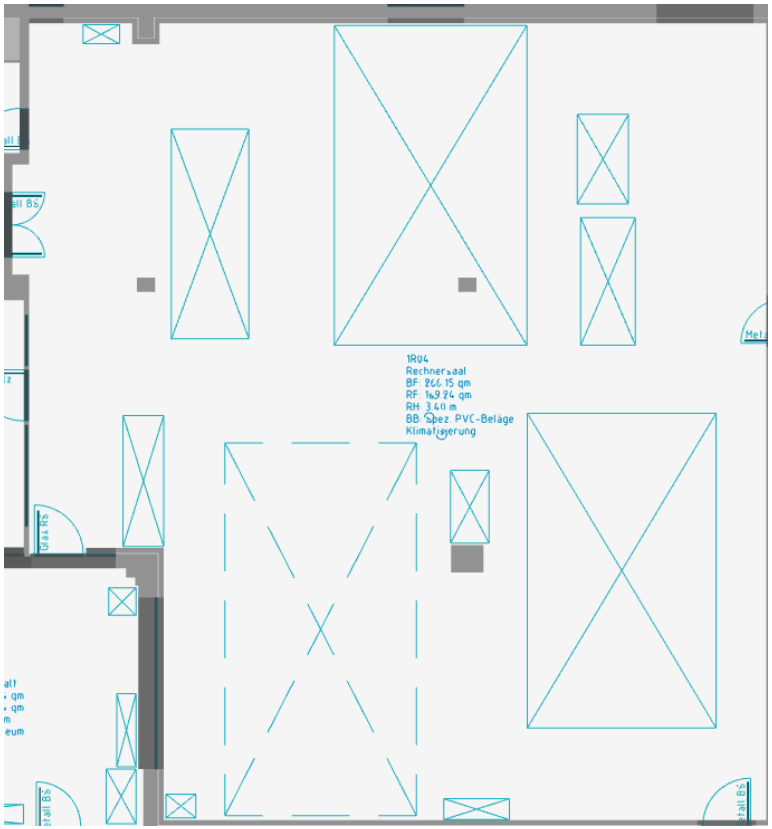
Backup

Die Räume des Computing Center

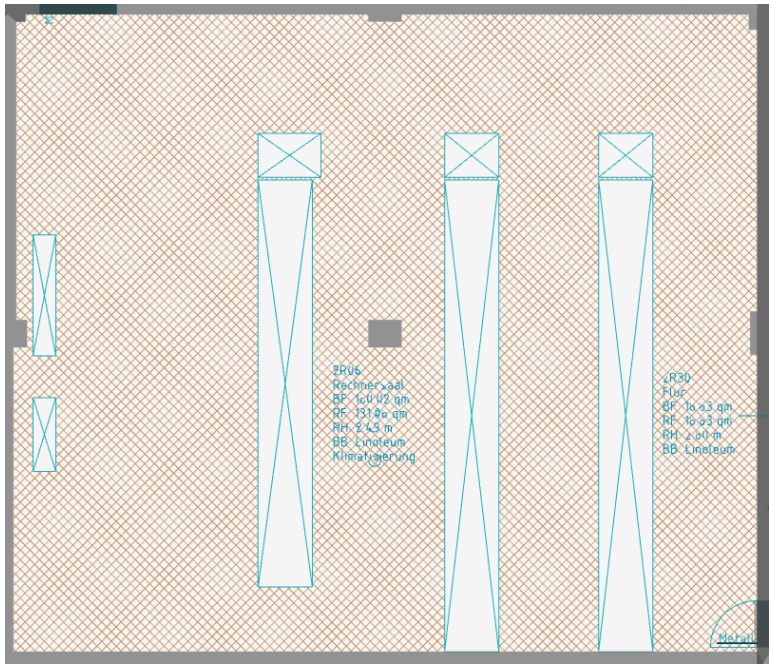
Infrastrukturen

verfügbar 426 qm beiten Platz für 76 Racks, 500kW Kühlung, 600kW USV

genutzt 300 qm mit 52 Racks, 200kW Kühlung, 200kW USV



EG - 1R04



OG - 2R06

Overview

Hardware

- 16,5 PB Disk, 7 PB Tape
- 10.000 Kerne
- 120 GPUs
- 700 Server
- 4 x 10 Gb/s Verbindung nach HH
- Extra 8 Gb/s Gäste Netz
- Firewall mit 30 Gb/s

Personal: 28+4

- 7 User/Office Support
- 5 Netzwerk
- 8 Software Entwicklung
- 8 Scientific Computing
- 4 Ausbildung

Services

- Self developed asset management system (*)
- Ansible configuration management (*)
- RHEL based operating systems (*)
- afs home directories (*)
- dCache mass storage with tape back end management software (*)
- Lustre scratch space
- Ceph storage for home dirs, VMs, S3 storage, samba (*)
- MinIO S3 Backup
- HTCondor for Grid and Batch farm
- Slurm parallel cluster
- Kubernetes cluster for container (*)
- CI/CD pipelines via Jenkins, gitlab
  - with harbor, sonarqube, nexus
- Monitoring and alerting via loki/prometheus/grafana/open search stack and LibreNMS
- MariaDB and MongoDB databases, Influx DB, MQTT Service
- jupyter in testing (*)
- Redmine
- Backup, dns, dhcp, . . .
- Experiment related services: RNO-G, ILDG, . . .
- Firewall protecting internal and DMZ networks
- Windows Server Environment

Nachhaltigkeit - Energie Effizienz Gesetz

Anforderungen des Gesetzgebers

- 2027: Power Usage Effectiveness unter 1.5
  - PUE ist der Gesamtverbrauch geteilt durch die Verbrauch der Server und Netzwerkkomponenten
  - Wir sind bei 1,27 bis 1.36 im Umbau
  - ab 2030: unter 1.3 wird gefordert
- 10% der Abwärme muss genutzt werden (nur bei neuen Rechenzentren)
- 2027: 100% erneuerbare Stromquellen
- 2026: EMS und UMS muss validiert werden

- Wir versuchen schon immer effizient zu sein, ökonomisch und ökologische, darum . . .
- . . . sind wir bereit (im Augenblick zumindest . . .)
- Natürlich arbeiten wir weiter daran, besser zu werden und die Technologie sinnvoll einzusetzen