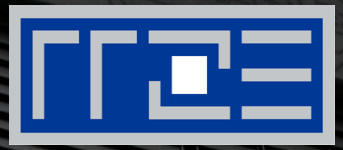
ERLANGEN REGIONAL
COMPUTING CENTER [RRZE]



HPC-Bedarf und HPC-Strategie 2020

RRZE-Campustreffen, 18.06.2015

Dr. Thomas Zeiser / Prof. Dr. Gerhard Wellein, RRZE

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PART 1: OPERATIONAL TOPICS

Coupling of HPC to IdM

- HPC systems will soon be provisioned through IdM

→**No HPC account without a valid IdM affiliation!**

→Expiration date of the HPC account may be shorter than requested if there is no IdM affiliation with long enough duration.

Long term plans: (details are not fixed yet)

- Work flow process to request HPC access instead of papers.
- Automatic prolongation of HPC accounts of *staff members* as their employment gets extended (up to the duration of the HPC project)

Expired / orphaned HPC accounts

- Do not expect that RRZE will keep data of expired HPC accounts for ever.
  IdM will make it easier to detect orphaned HPC accounts.
- Data may be purged 3 months after *expiration of the HPC permission*. (See back side of HPC form.)
  (**ATTENTION:** if your HPC account is identical to your IdM account!)

- If data shall be transferred to a different account we need the permission of the original owner.

Reminder: Export control & non-proliferation HPC systems and research are Dual Use goods

- cf. HPC-Kolloquium from October 2012
  http://www.rrze.fau.de/dienste/arbeiten-rechnen/hpc/kundenbereich/HPC-Koll_301012.pdf
- cf. letter of ZUV from 19.11.2014
  http://www.zuv.fau.de/universitaet/organisation/verwaltung/zuv/verwaltungshandbuch/drittmittel/Exportkontolle_bei_Forschungsleistungen_-_BAFA_-_Au%C3%9Fenwirtschaftsgesetz.pdf

Official legal information is only available from *Bundesamt für Wirtschaft und Ausfuhrkontrolle (BAFA)*

- http://www.ausfuhrkontrolle.info/ausfuhrkontrolle/de

Some readable notes:

- http://www.bmbf.de/pub/supercomputer_und_exportkontrolle.pdf
- http://www.bafa.de/ausfuhrkontrolle/de/arbeitshilfen/merkblaetter/merkblatt_tt.pdf

HPC systems @ RRZE
hardware overview / news

- **Woody:** single node (throughput) jobs
  - The last old w0xxx nodes have been switched off in 09/2014
  - w10xx (48) and w11xx (72) have modern CPUs but still 8 GB/node
- **LiMa:** nodes are slowly dying due to cooling failure in 06/2014
  - Currently approx. 440 out of originally 500 nodes available
- **Emmy:** business as usual – sometimes quite long queue times
- **TinyGPU:** GPUs of the original nodes (tg0xx) no longer supported by latest NVidia drivers; use tg0xx for non-GPU load.
- **TinyBlue, TinyFAT, Windows cluster, HPC storage:** no news

HPC systems @ RRZE
software news / plans

- **TinyGPU/TinyBlue/TinyFAT:**
  - OS upgraded from Ubuntu 12.04 to 14.04 (already in Feb./Apr.)
  - Use “woody3.rrze” as Ubuntu front end
- **Woody:**
  - Still running SuSE SLES 11SP3
  - Might be reinstalled with Ubuntu 14.04 in the future.
- **LiMa/Emmy:**
  - OS upgrade from CentOS 6.x to 7.x planned for later this year.

New HPC systems beyond Erlangen

- Inauguration of **SuperMUC phase-2** will be in June/July
- Access requires scientific proposal:
  https://www.lrz.de/services/compute/supermuc/projectproposal/

- **LRZ's new Linux cluster** will soon be operational
  - Will consist of hardware similar to half a SuperMUC phase-2 island.
  - Accounts can easily be requested through RRZE.

Next HPC system for Erlangen

- RRZE presented a *Forschungsgroßgeräteantrag* at FAU's KORA.
- There was a long discussion whether FAU can afford a new HPC cluster every 3 years – but the proposal finally passed.
- It took DFG three months to acknowledge receipt of the proposal.
- No idea how long evaluation will take …

- The new system will be installed at the same position as LiMa.

  → LiMa must be disassembled before the new system can be brought in.

  → Expected size: at most as large as Emmy – probably (slightly) smaller.

How expensive are the current HPC systems @ FAU?

- **Hardware:** a new cluster (~2,5 Mio EUR) every 3 years

  ➔ less then 1 Mio EUR/year
- **Running costs for electricity and cooling**
  - Average power input of all current HPC systems: >300 kW * 365x24
  - Cooling efforts: PUE > 2.0 (could be reduced but infrastructure work is pending)

  ➔ almost 1 Mio EUR/year

  (electricity and cold water are just available from the wall socket and nobody cares [yet])
- **HPC staff at RRZE** (<2,5 FTE)

  ➔ less then 200 kEUR/year

https://www.zuv.fau.de/universitaet/organisation/verwaltung/zuv/verwaltungshandbuch/haushalt/FAU_Haushaltsplan.pdf

15 19 Universität Erlangen-Nürnberg

Titel	FKZ	Zweckbestimmung	2015 Tsd. €	2016 Tsd. €	A B C	Soll 2014 Ist 2013 Ist 2012 Tsd. €
1	2	3	4	5		6
		76 Einrichtung und Ausstattung neuer, sowie Ergänzung der Einrichtung und Ausstattung bestehender Hochschuleinrichtungen *Die Ausgabebefugnis erhöht sich um die Isteinnahme bei 331 04 und 331 07.* *Vgl. Vermerk zu 519 01, 812 01, TG 73 und zu Kap. 15 28 TG 75.*				
812 76-5	133	Erwerb von Geräten, Ausstattungs- und Ausrüstungsgegenständen im Inland	3.216,7	3.314,1	A B C	3.000,0 9.678,5 8.198,7
		Summe der Titelgruppe	3.216,7	3.314,1	A B C	3.000,0 9.678,5 8.198,7
517 05-9	133	Bewirtschaftung durch Heizung, Beleuchtung und elektrische Kraft	18.786,6	18.938,1	A B C	18.010,3 18.448,7 16.147,6

PART 2: “WHAT NEXT?”

What next: Clusters at RRZE: 2003 – 2013

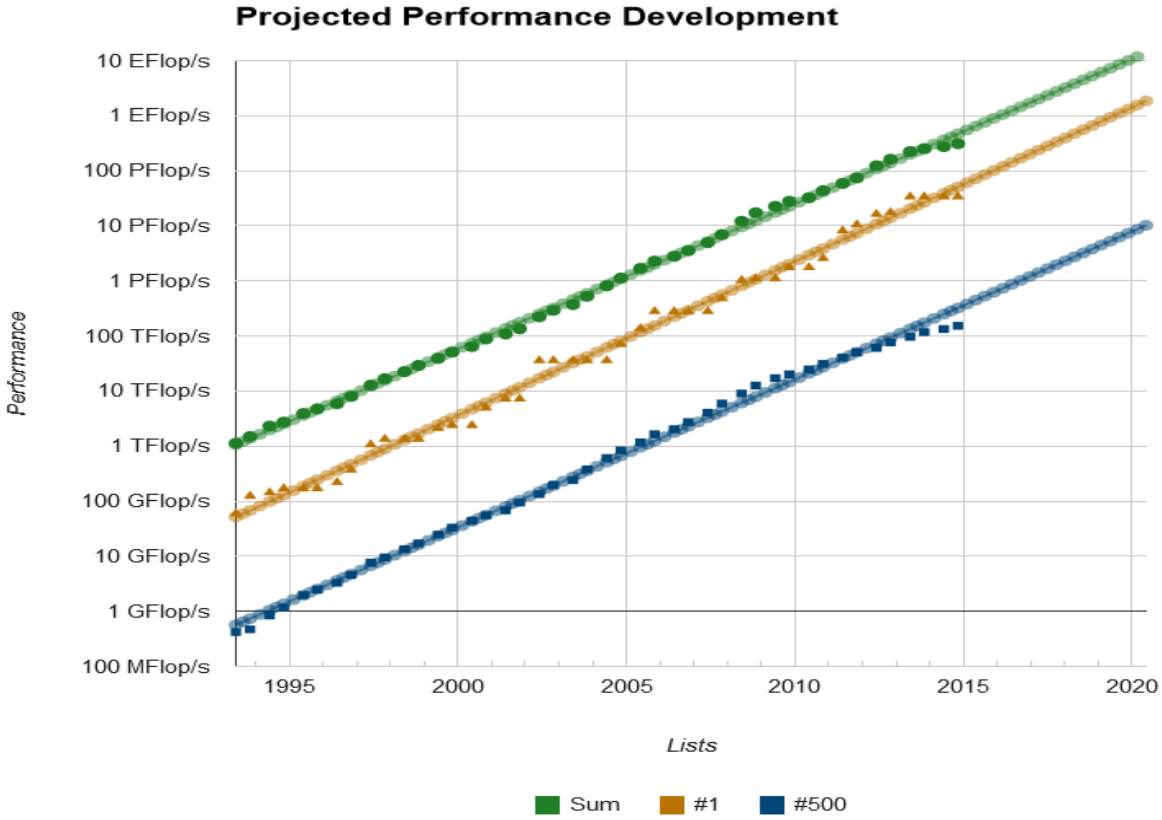
- RRZE: Install a new HPC cluster every 3 years → Art. 91b
- At least two systems are operated concurrently
- Between 20% and 50% of investment cost contributed directly by scientists (“Rezentralisierung”)

What next: Clusters at RRZE: 2003 – 2013

Node specs	#Nodes	#Cores	Price	Peak	TOP500	Year
2 x Intel Xeon 2.66 GHz; 2 GB	77	154	0.35 M€	0.8 TFlop/s	315	2003
2 x Intel Xeon 3.0 GHz; 8 GB	182	728	1.0 M€	8.7 TFlop/s	124	2006
2 x Intel Xeon 2.66 GHz; 24 GB	500	6.000	2.3 M€	64 TFlop/s	130	2010
2 x Intel Xeon 2.2 GHz; 64 GB	560	11.200	2.6 M€	234 Tflop/s	210	2013

- Proposal for new system sent to DFG (2.5 M€)

TOP500 – still looking good?



Budget increase –
RRZE x86 Cluster:

2003 → 2010: 6.0x
2010 → 2013: 1.1x
2013 → 2016: 0.96x

What next?! Clusters at RRZE: 2003 – 2013

Trends 2003 – 2013:

- Price per node (including all infrastructure): ~ 4.500-5.000 €
- Power consumption per node (CPU only) ~ 300-350 W
- Power was not an issue for RRZE in the past (HPC systems including cooling contributed less than 5% of “Technische Fakultät” campus
  (“10 MW power line”))
- Cluster nodes TOP120-130 (Nov. 2012):
  ~660 nodes 8c-SNB (AVX vs. SSE)

What next?! Clusters at RRZE: 2003 – 2013

TOP120-130 installation in 2016: ~1.500 nodes

- Space: 20-30 Racks
- Power consumption system only (**PUE=2**): 0.5 MW (1 MW)
- Operating this machine is not feasible with **existing** infrastructure & concept

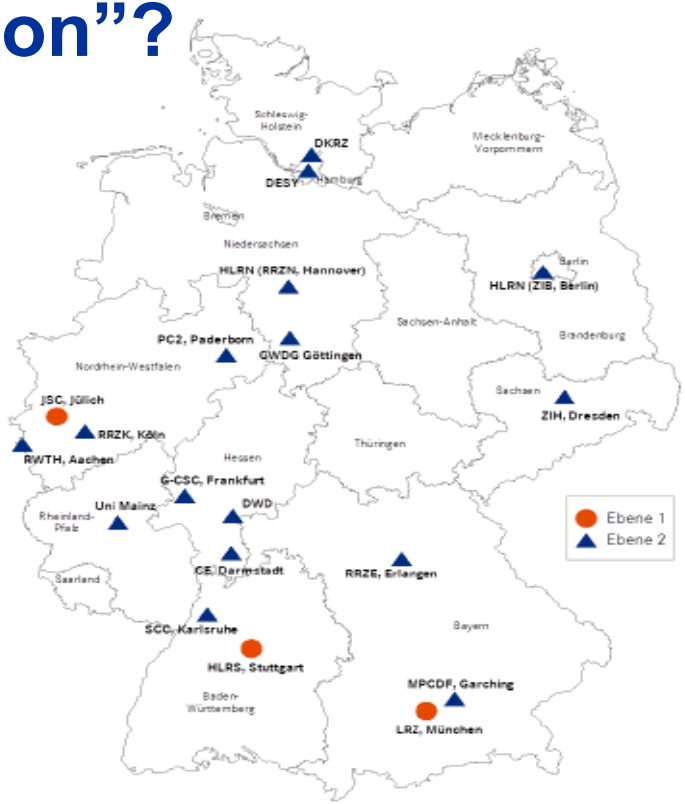
→ RRZE will not increase node counts in near future

→ Less competitive compute resources for science and research at FAU

What's up with the "competition"?

Construction of new data centers / computer rooms (since 2000):

- RWTH Aachen
- TU Darmstadt
- TU Dresden
- Uni. Köln
- Paderborn, Siegen,…
- Big 3: Stuttgart, Jülich, München



Quelle. Wissenschaftsrat; Standorte wg. Lesbarkeit tw. angepasst. Kartengrundlage © Lutum+Tappe

What's up with the hardware?

	SIMD	ILP	Cores*Clock (RRZE) [cores x GHz]
Woodcrest	128 Bit	MULT + ADD	6
Westmere	128 Bit	MULT + ADD	**16**
IvyBridge	**256 Bit**	MULT + ADD	22
Haswell	256 Bit	**FMA + FMA**	32

Jugglers' tricks of computer architects: SIMD & FMA

Basic limitations:

- SIMD → 512 Bit max.
- ILP → 4 FMAs useless?! (instruction throughput: 4 Instr./cyce)
- Cores*Clock → Next slide

→ Accelerator?! (particulary good juggler)

What's up with the hardware?

	Socket Config.	Socket-Speed	TDP	Price [USD]
Nehalem	**4 cores*3.33 GHz**	**13.3 core*GHz**	**130 W**	**1600**
Westmere	6 cores*3.46 GHz	20.8 core*GHz	130 W	1663
Sandy Bridge	**8 cores*3.1 GHz**	**24.8 core*GHz**	**150 W**	**1885**
Ivy Bridge	12 cores*2.7 GHz	32.4 core*GHz	130 W	2614
Haswell	**18 cores*2.3 GHz**	**41.4 core*GHz**	**145 W**	**> 2700**

Data: http://en.wikipedia.org/wiki/ → Look for specific microarchitecture

Top Bin for 2-way EP servers

Cores*GHz slows down ?!

Price increases!

Accelerators: Trade in code flexibility/quality for performance

“Status Quo” RRZE

Current status: RRZE freezes system size on LIMA + EMMY level

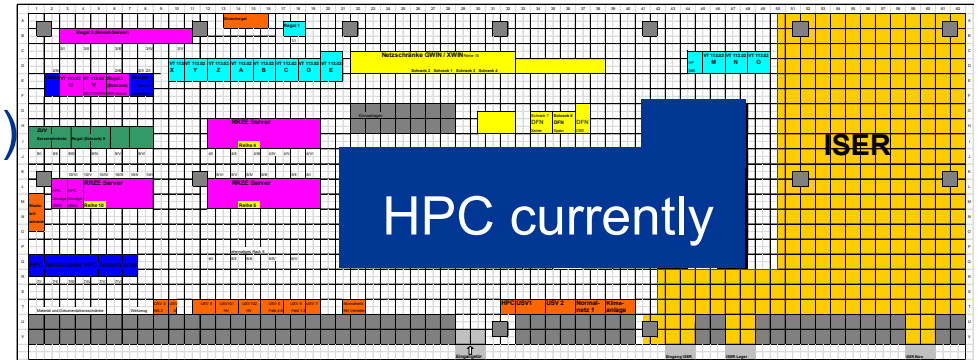
→ Aiming at a procurement of a 500 nodes system every 3 years
→ 2 systems are operated simultaneously (+HPC Storage)

→ Qualitative growth of computing power no longer possible
(beyond Socket-Speed / Accelerator)

There will be a detailed survey of the HPC needs by ZISC soon!

Alternatives on acting

- Computing time in Erlangen is sufficient
- Applying for compute time in München/Stuttgart/Jülich/GCS/PRACE/…
- ISER “Acquisition” + Update
  of the infrastructure (>1 Mio EUR)
  → Potential: nodes x 2
  but also 2x electricity and HW!



- Reconstruction HPC server room or complete “RRZE”: timeline unknown
  RRZE-initiatives since 2010 have been unsuccessful / without progress

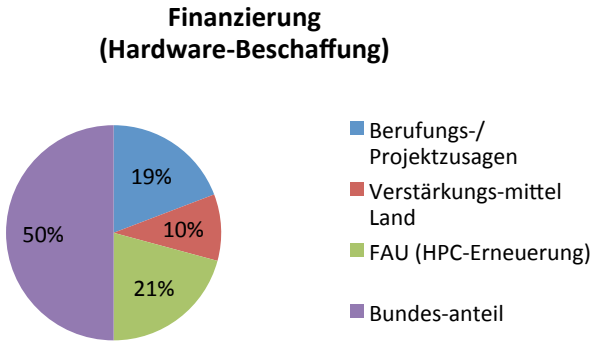
ANLAGE:

Ausgewählte Folien aus der KoRa-Sitzung am 9. Feb. 2015 zur Vorstellung des Forschungsgroßgeräte-Antrags „Hochleistungscomputecluster“

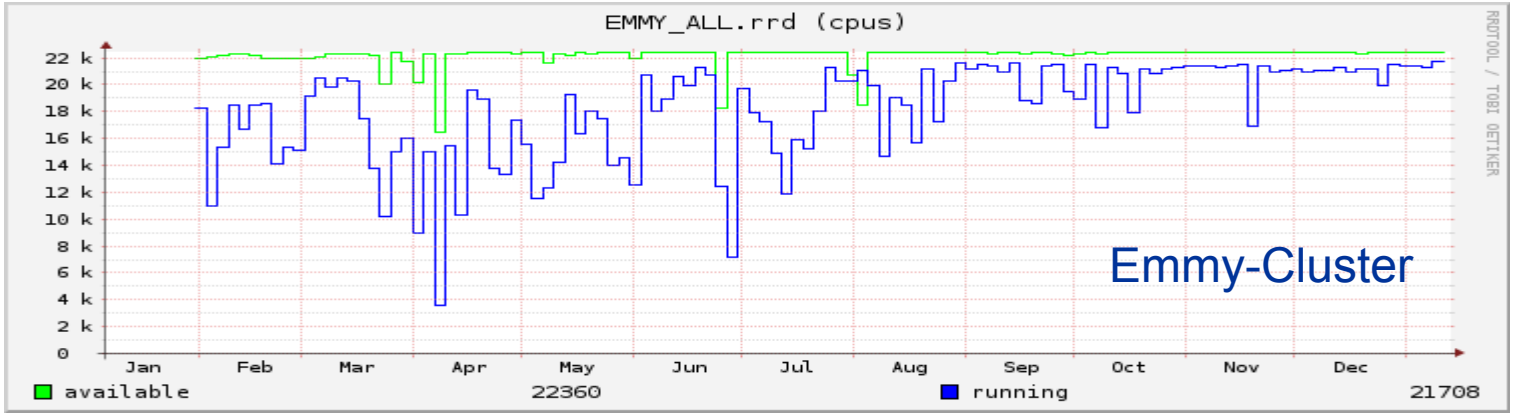
FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

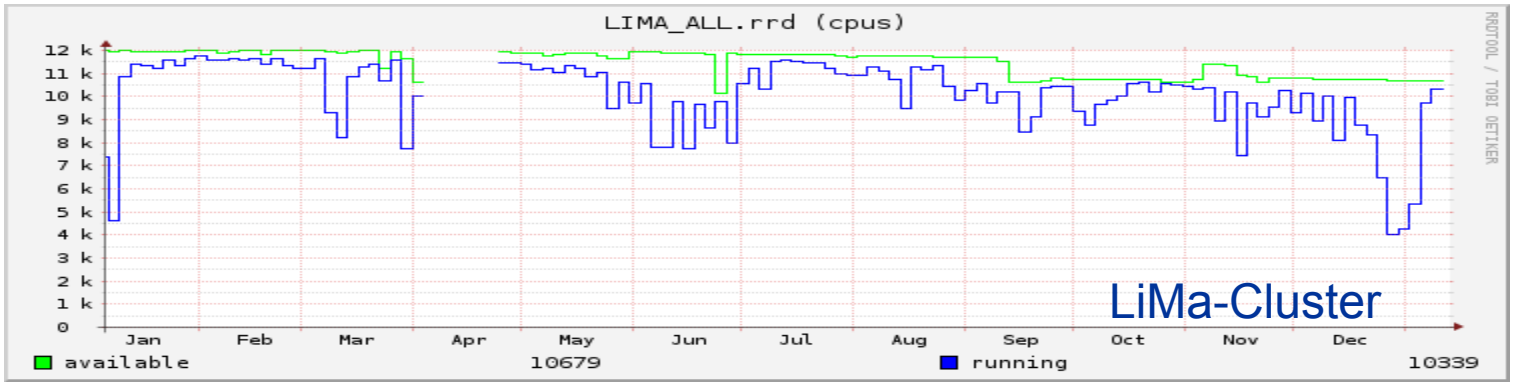
Beantragt wird …

- ein **Forschungsgroßgerät (bis 5 Mio €) nach Art. 91b GG**
- *„Neuer HPC-Cluster insbesondere für numerische Simulation in der Chemie und Biologie (Lebenswissenschaften) sowie den Material- und Ingenieurwissenschaften und der Geographie/ Klimatologie.“*
- ein **Gesamtbugdet von 2,5 Mio €**

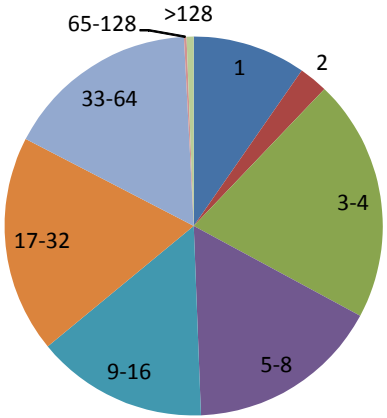


Auslastung der großen HPC-Systeme im Jahr 2014



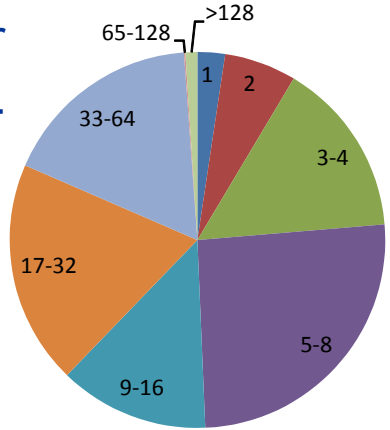


Jobgrößenverteilung auf Emmy und LiMa



Emmy-Cluster
Im Jahr 2014

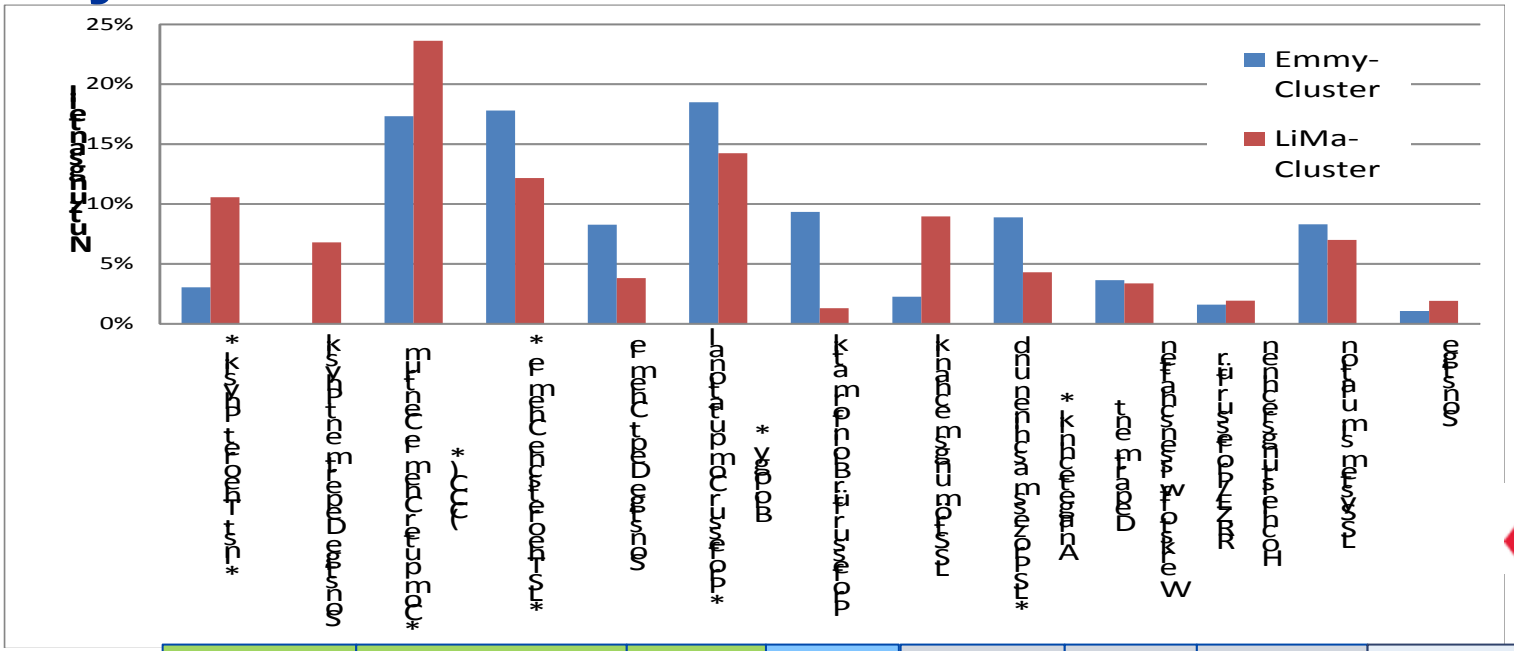
LiMa-Cluster
Im Jahr 2014



Über 50% der Rechenzeit entfällt auf parallele Jobs mit mindestens 9 Knoten (d.h. mindestens 180 bzw. 108 Kernen).

➔ Parallelrechner mit gutem internen Netzwerk ist zwingend notwendig

Nutzungsanteil der wichtigsten Gruppen des Emmy und LiMa-Clusters im Jahr 2014



>50%

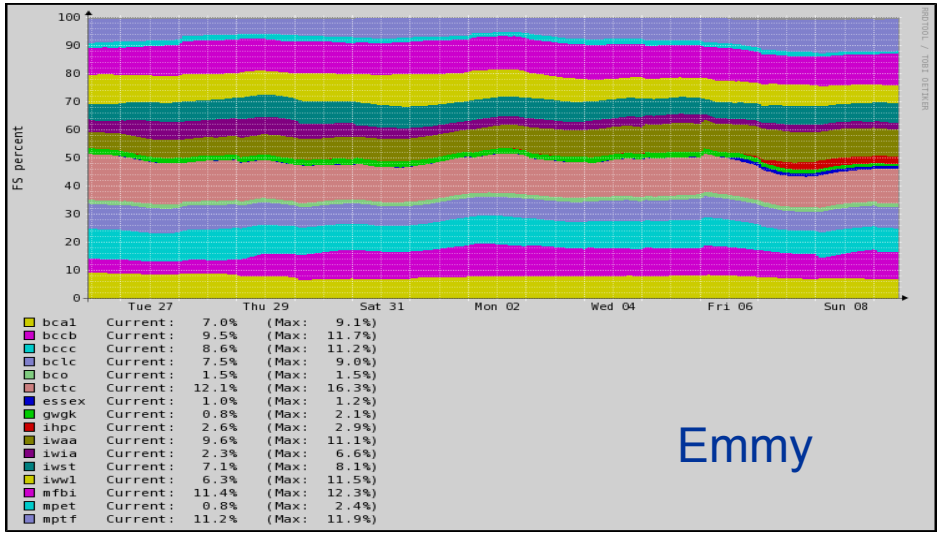
Gewichtet über alle Systeme	Physik 10%	Chemie 42%	Bio 17%	Med 6%	CBI 11%	WW 3%	Inf 9%	Rest 2%

Detaillierter Rechenzeitverbrauch im Jahr 2014

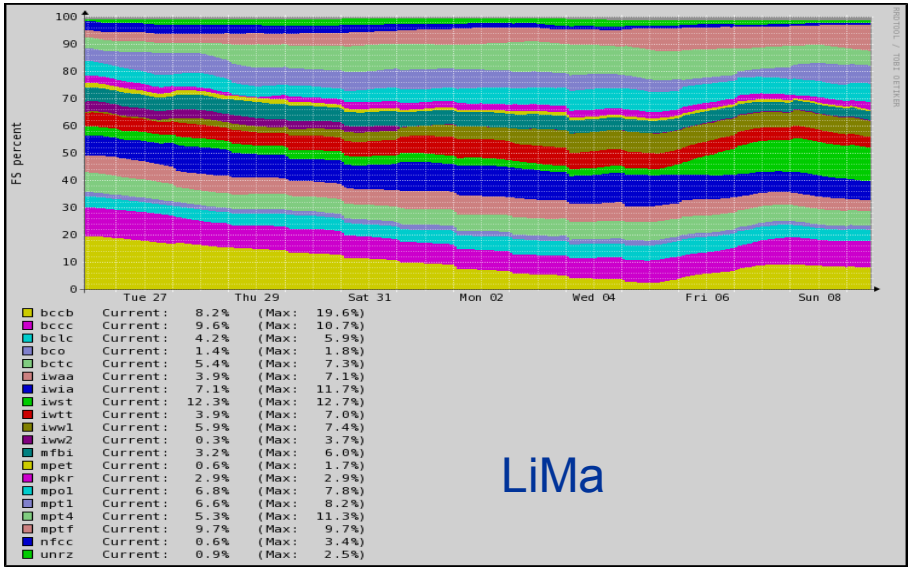
Cluster		Emmy	LiMa	Woody	TinyBlue	TinyFAT	TinyGPU	Windows	gemäß den im Jahr 2014 geltenden RRZE-Preislisten entspricht die insgesamt abgegebene Re enzeit einem W er 5,3 Mio C osten wie n)	
insgesamt abgegebene Rechenzeit (in SMT-Core-Stund		156.242.374	87.129.678	4.220.121	7.177.648	520.680	172.092	381.557		
insgesamt abgegebene Rechenzeit (in Knoten-Stunden)		3.906.059	3.630.403	1.055.030	448.603	32.543	10.756	31.796		
Auslastung		*84%*	*88%*	*81%*	*63%*	*25%*	*14%*	*23%*		
Wert der Rechenzeit gemäß RRZE-Preisliste		3.124.847 €	1.815.202 €	211.006 €	179.441 €	19.526 €	4.302 €	19.07		
Nutzungsanteil										
NatFak										69,0%
	Erlangen Centre Acroparticle Physics (ECAP)			19%	<1%	2%			,3%	
	Inst. Theoret. Physik	3%	11%	12%	17%	3			6,5%	
	Sonstige Depertment Physik		7%	1%	<1%				2,3%	
	Computer Chemie Centrum (CCC)	17%	24%	22%	80%				21,6%	
	LS Theoretische Chemie	18%	12%	1%	1				14,7%	
	Sonstige Dept. Chemie	8%	4%	1%					6,1%	
	Professur Computational Biology	18%	14%	29%					16,7%	
	Sonstige Dept. Biologie	0%	<1%						0,0%	
	Dept. Geographie								0,0%	
	Dept. Mathematik		<1%		1%				0,3%	
MedFak										6,1%
	Professur für Bioinformatik	9%				3%	67%		6,0%	
	Sonstige Med. Fakultät				1%				0,1%	
TechFak										24,0%
	LSTM					<1%			4,5%	
	IPAT			2%	1%	2%	4%		6,7%	
	Sonstige Dept. CBI			1%	<1%	<1%	1%		0,2%	
	Dept. EEI								0,0%	
	Dept. WW		3%	2%		<1%			3,3%	
	Dept. MB		<1%	1%			8%		0,0%	
	RRZE/Professur für H		2%	<1%	<1%	<1%	2%		1,6%	
	LS Systemsimul	8%	7%				<1%		7,2%	
	Sonstige	<1%	<1%	4%	<1%	<1%	<1%		0,4%	
Wirtschaftswi										0,4%
								58%	0,2%	
								42%	0,1%	
								<1%	0,0%	
P										0,0%
	nguistik				<1%	1%			0,0%	
Ext	iges									0,5%
	gsaufzeichnung / Videokodierung					9%			0,0%	
	terne Projektpartner	1%	<1%	2%	<1%		1%		0,5%	
	Uni Bamberg, HS Coburg+Nürnberg		<1%						0,0%	
Cluster		Emmy	LiMa	Woody	TinyBlue	TinyFAT	TinyGPU	Windows		

Siehe jährlichen RRZE-Jahresbericht!

Fairshare-Verteilung der letzten 10 Tage (Feb. 2014)

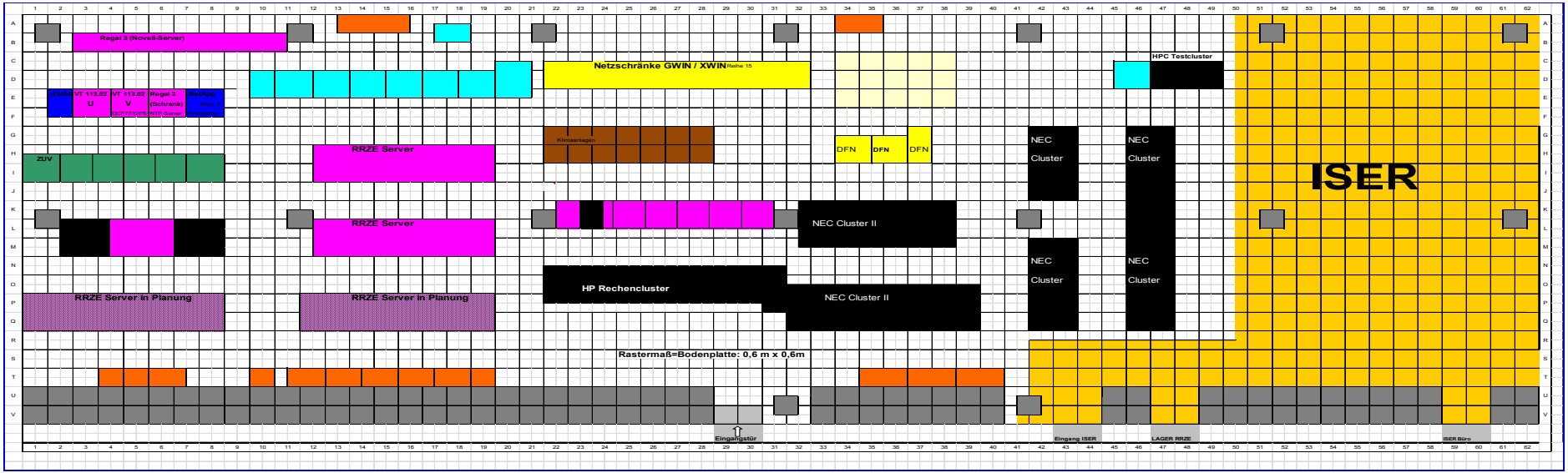


Emmy



LiMa

Stellplan RRZE-Rechnerraum



	Server (RRZE-Server, Hosting, Housing)
	ZUV-Server
	HPC-Systeme
	DFN-Infrastruktur
	Netz-Komponenten (RRZE)
	Sicherungsschränke / Stromverteiler
	Raum-Klimageräte

Aufstellungsort für beantragtes System

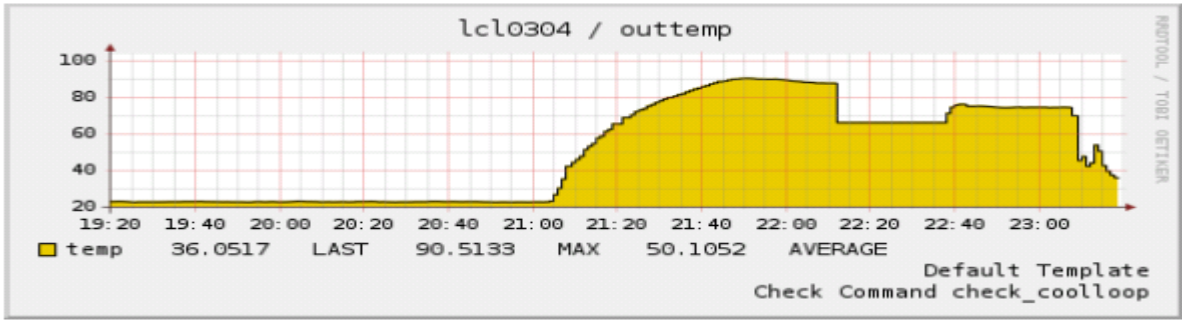
- LiMa-Cluster (2010) und TinyFAT (2010) raus
- Neues System rein
- Für mindestens 2 Monate wird den Wissenschaftlern maximal 2/3 der jetzigen Rechenleistung zur Verfügung stehen

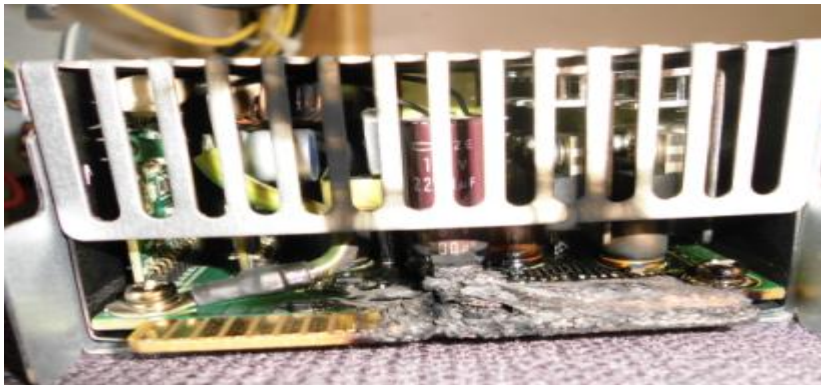
- Zum angepeilten Installationszeitpunkt wird LiMa >5 Jahre alt sein
- Abschreibung gemäß DFG-Richtlinien über 4 Jahre

Aufstellungsort für beantragtes System

- Durch Ausfall der Kaltwasserversorgung in der Nacht vom 24. Juni ist es zu massiver Überhitzung von LiMa gekommen
  → unmittelbare Hardware-Ausfälle
  → massiv beschleunigte Alterung → Ausfälle / Instabilitäten
- Stand heute:
  bereits 53 von 500 Rechenknoten nicht mehr nutzbar

Ausfall der Kaltwasserversorgung in der Nacht vom 24. Juni 2014 & Folgen





Frühzeitiger Cluster-Tausch lohnt sich finanziell für die Uni: Beispiel Woody-Cluster

- Anschaffung/Erweiterung: 2006/2007 (ca. 1,5 Mio EUR)
- 225 Rechenknoten mit einer Stromaufnahme von >75 kW
  + Energieaufwand für die Kühlung

➔ Stromkosten 1 Mio kWh/a = 150k €/a

Frühzeitiger Cluster-Tausch lohnt sich finanziell für die Uni: Beispiel Woody-Cluster

- Abschaltung von Woody in 2013/2014 obwohl Hardware noch lief
- Ersatz durch 6 moderne, preisoptimierte Enclosure mit je zwölf 1-Socket-Systeme für seriellen Durchsatz
  Anschaffungspreis: 60k EUR (30k EUR durch FAU „gesponsort“)
- Die 72 Knoten bringen für Durchsatzjobs etwa die gleiche aggregierte Rechenleistung wir die alten 225 Rechenknoten
- Stromaufnahme < 8 kW + Energieaufwand für die Kühlung

➔ Stromkosten 0,1 Mio kWh/a = 15k EUR/a

HPC-Systeme am RRZE: gestern, heute, morgen

HPC storage
(60 TB disk)

2009

throughput cluster

2003

parallel cluster(s)

2006/2009

"fat" nodes
up to 512 GB mem

2001 / 2003 / 2010

HPC storage
HSM to tape

2009

10 Gbit HPC-Ethernet
backbone

small research cluster
with GPUs
2009 / 2010 / 2013

2010

high-end parallel cluster
130@Top500 11/2010 (2,0 Mio €)

2013

210@Top500 11/2013 (2,5 Mio €)

2008

small cluster
running Windows HPC

2009

Clusters@RRZE: 2003 – 2014

- A new HPC cluster every 3 years → Art. 91b
- At least two systems are operated concurrently
- 20% - 50% of investment costs contributed by scientists

Clusters@RRZE: 2003 – 2014

Node specs	#Nodes	#Cores	Price	Peak	TOP500	Year
2 x Intel Xeon 2.66 GHz; 2 GB	77	154	0.35 M€	0.8 TFlop/s	315	2003
2 x Intel Xeon 3.0 GHz; 8 GB	182	728	1.0 M€	8.7 TFlop/s	124	2006
2 x Intel Xeon 2.66 GHz; 24 GB	500	6.000	2.0 M€ +0.3 M€	64 TFlop/s	130	2010
2 x Intel Xeon 2.2 GHz; 64 GB	560	11.200	2.6 M€	234 TFlop/s	210	2013
2 x Intel Xeon 2.5 GHz; 64 GB	~300 -400	~6.000 -8.000	2.5 M€	~200-300 TFlop/s	n/a	2016

Thank you for your attention!

Regionales RechenZentrum Erlangen [RRZE]
Martensstraße 1, 91058 Erlangen
http://www.rrze.fau.de