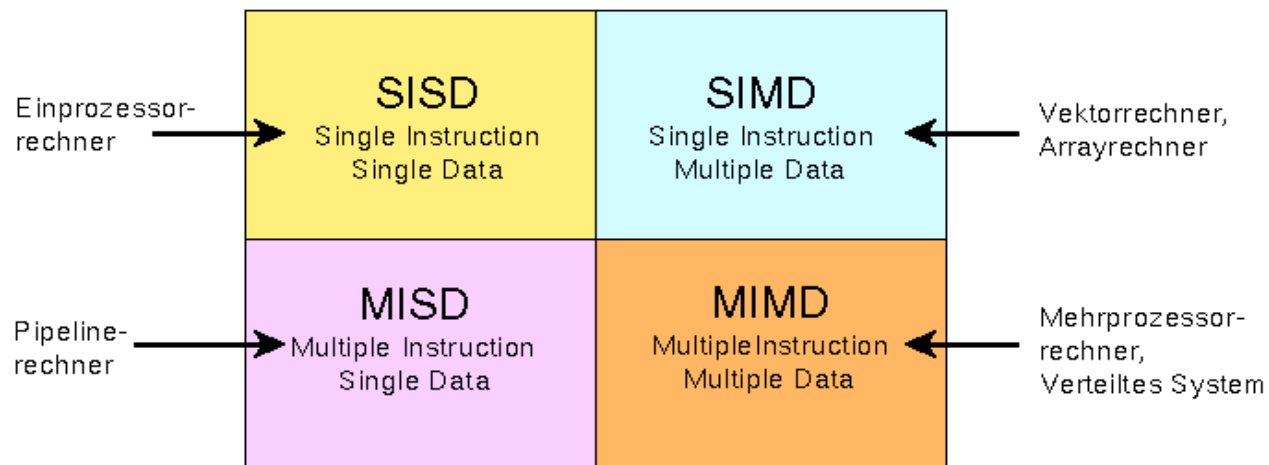




Über den Mainframe hinaus – Trends im Rechenzentrum

Prof. Dr. Andreas Polze
Operating Systems and Middleware, Hasso Plattner Institute
Faculty of Digital Engineering

Rechnerklassifikation



M.Flynn, *Very High Speed Computing Systems*,
Proceedings of the IEEE, vol 54, 1966, pp. 1901-1909(9)

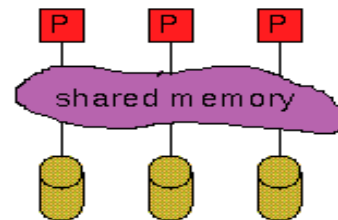
Über den
Mainframe
hinaus...

Andreas Polze

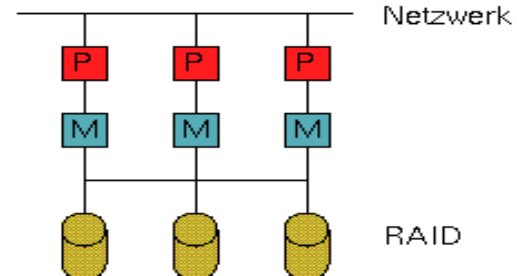
Chart 2

MIMD: heute Standard (Multicore/Manycore)

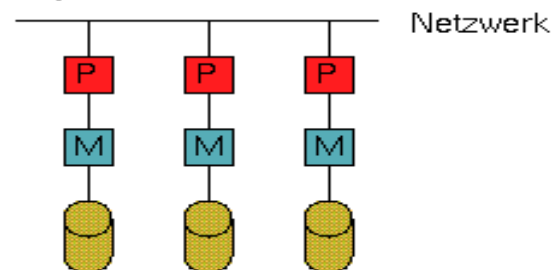
shared memory (enge Kopplung)
Mehrprozessorrechner



shared disks



shared nothing (lose Kopplung)
verteiltes System

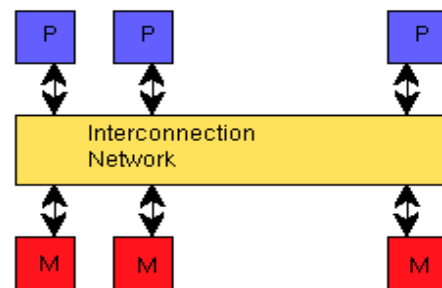


Über den
Mainframe
hinaus...

Andreas Polze

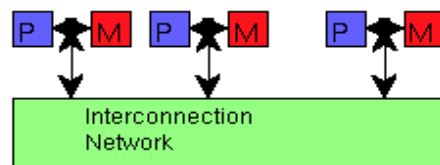
Chart 3

Hardware details schimmern durch...

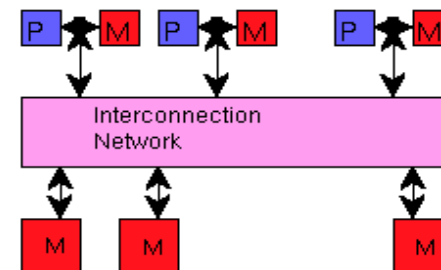


Uniform-memory-access
shared-address-space
computer (UMA)

Non-Uniform-memory-access
shared-address-space
computer with local and
global memories (NUMA)



Non-uniform-memory-access
shared-address-space
computer with
local memory only (NUMA)



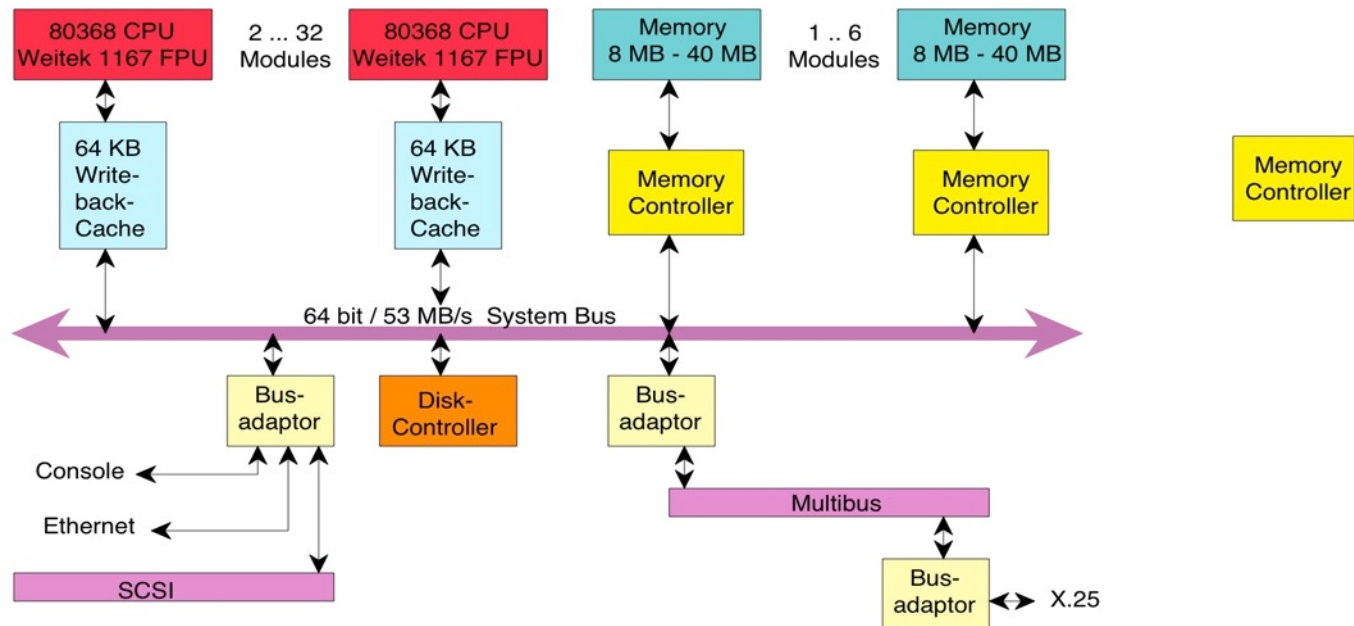
Über den
Mainframe
hinaus...

Andreas Polze

Chart 4

Klassische Beispiele

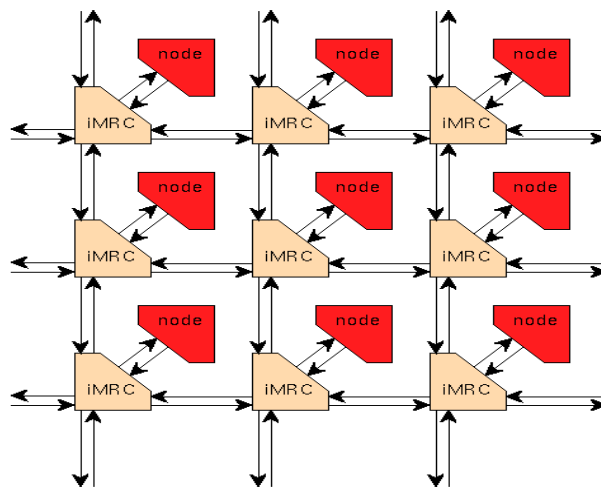
Sequent Symmetry



Über den Mainframe hinaus...
Andreas Polze

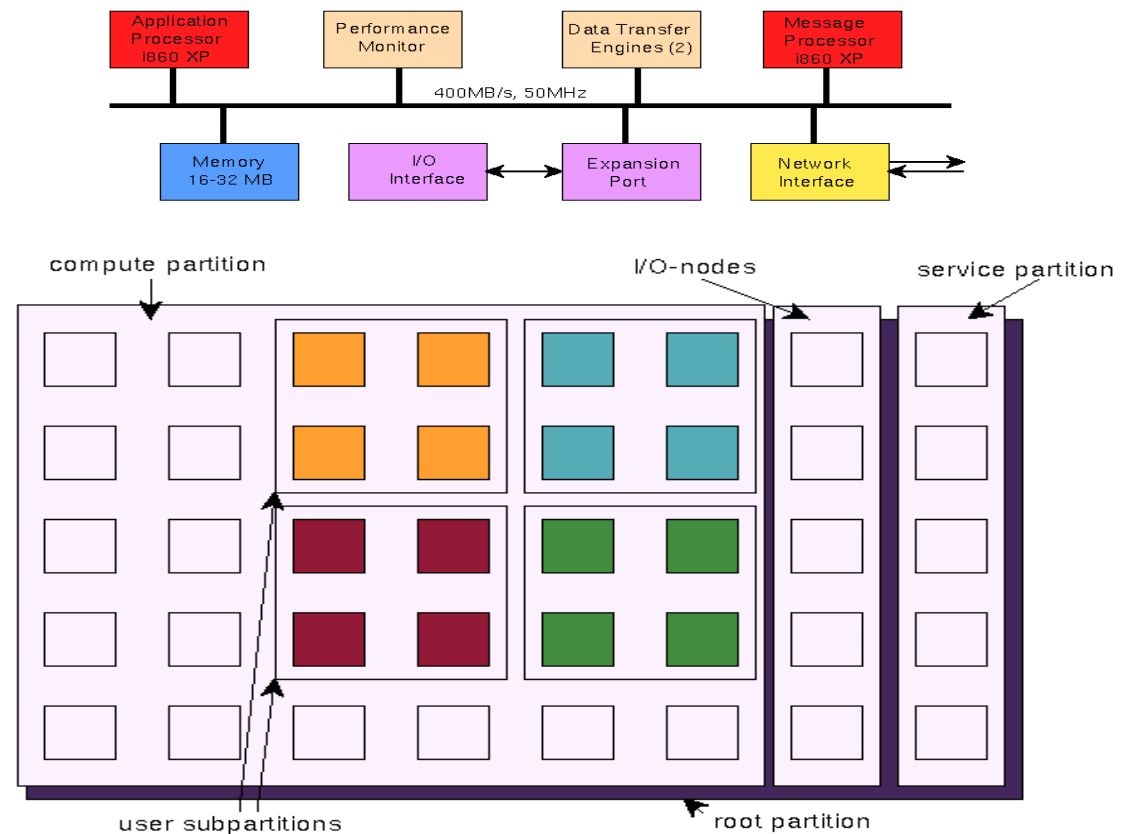
Chart 5

Intel Paragon



Meilenstein der
Rechnerarchitektur

Jeder Knoten mit
eigenem OS (Mach)

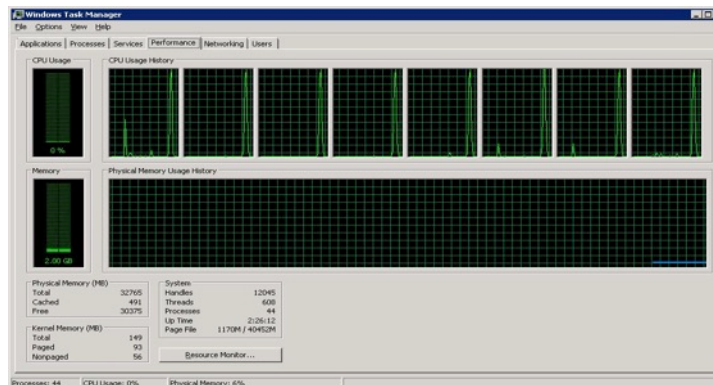
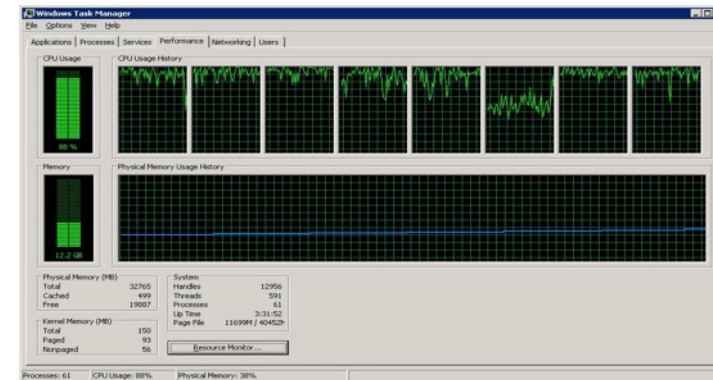


Standard-Server

Kompakt, Dicht
Standard Architekturen (x64)
Multicore/Multithreading

Standard-Betriebssysteme: Linux
Virtualisierung
Thrustworthiness/security
Clustering

Neue Programmiermodelle,
Software-Architekturen,
Services

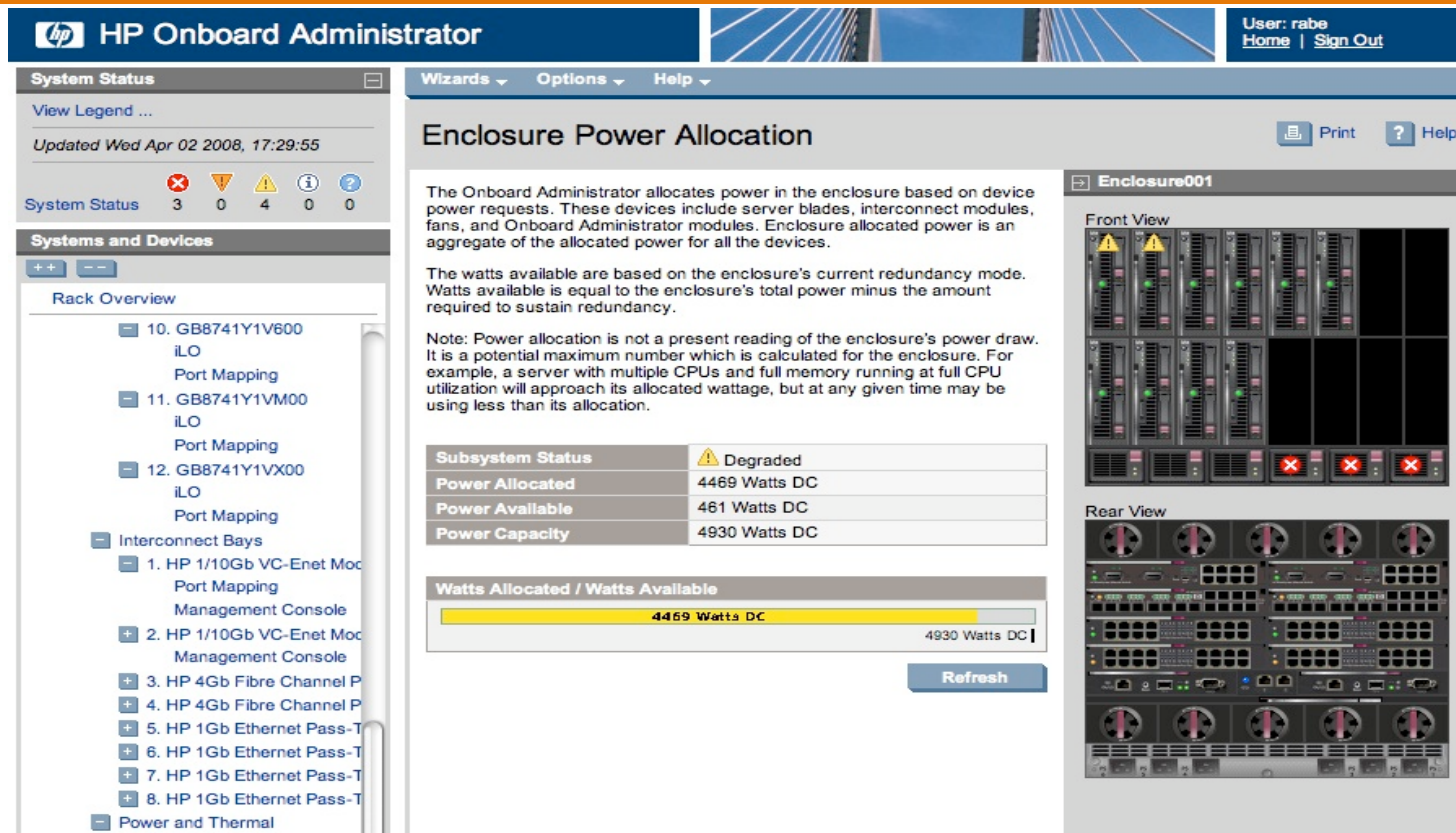


Über den
Mainframe
hinaus...

Andreas Polze

Chart 7

Energieverbrauch: Alte Probleme bleiben bestehen...



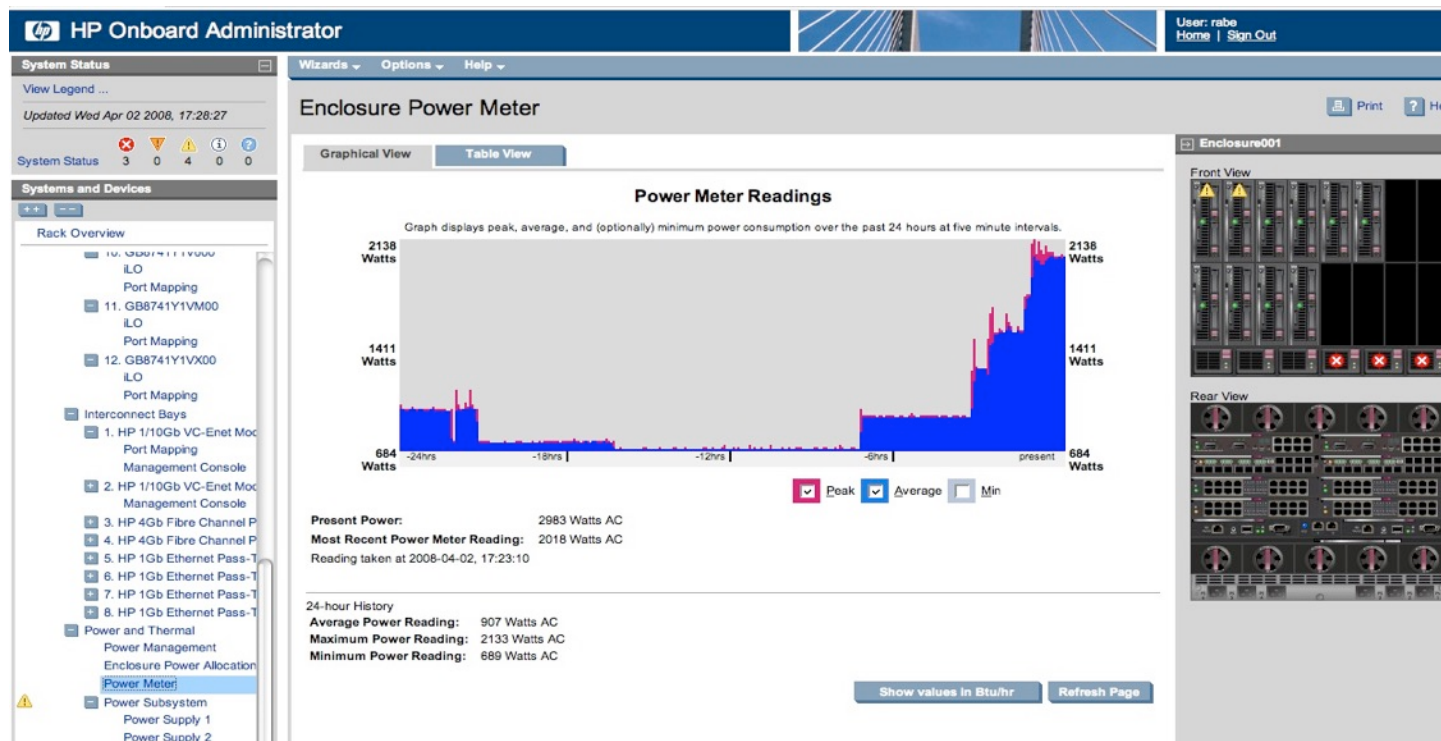
Über den Mainframe hinaus...

Andreas Polze

Chart 8

Green IT

Konsolidierung steigert Effizienz



Über den
Mainframe
hinaus...

Andreas Polze

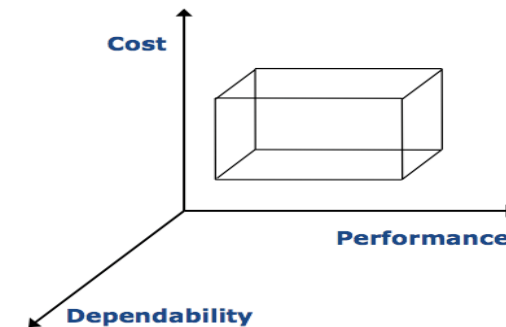
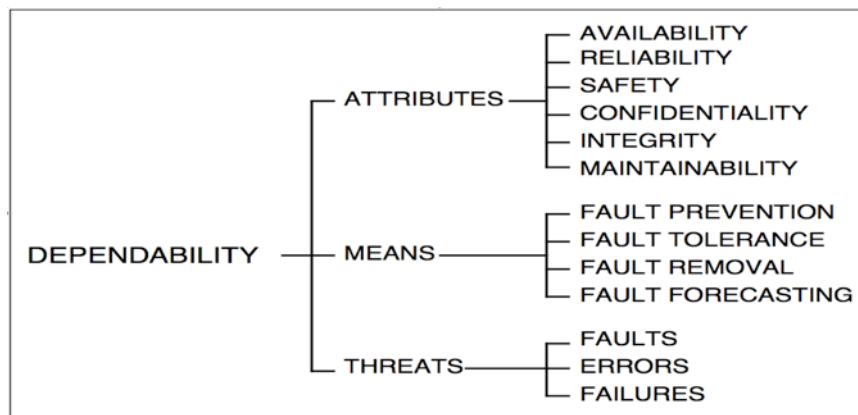
Chart 9

Dependability

Umbrella term for operational requirements on a system

- „Trustworthiness of a computer system such that reliance can be placed on the service it delivers to the user“ [Laprie]

General question: How to deal with unexpected events ?

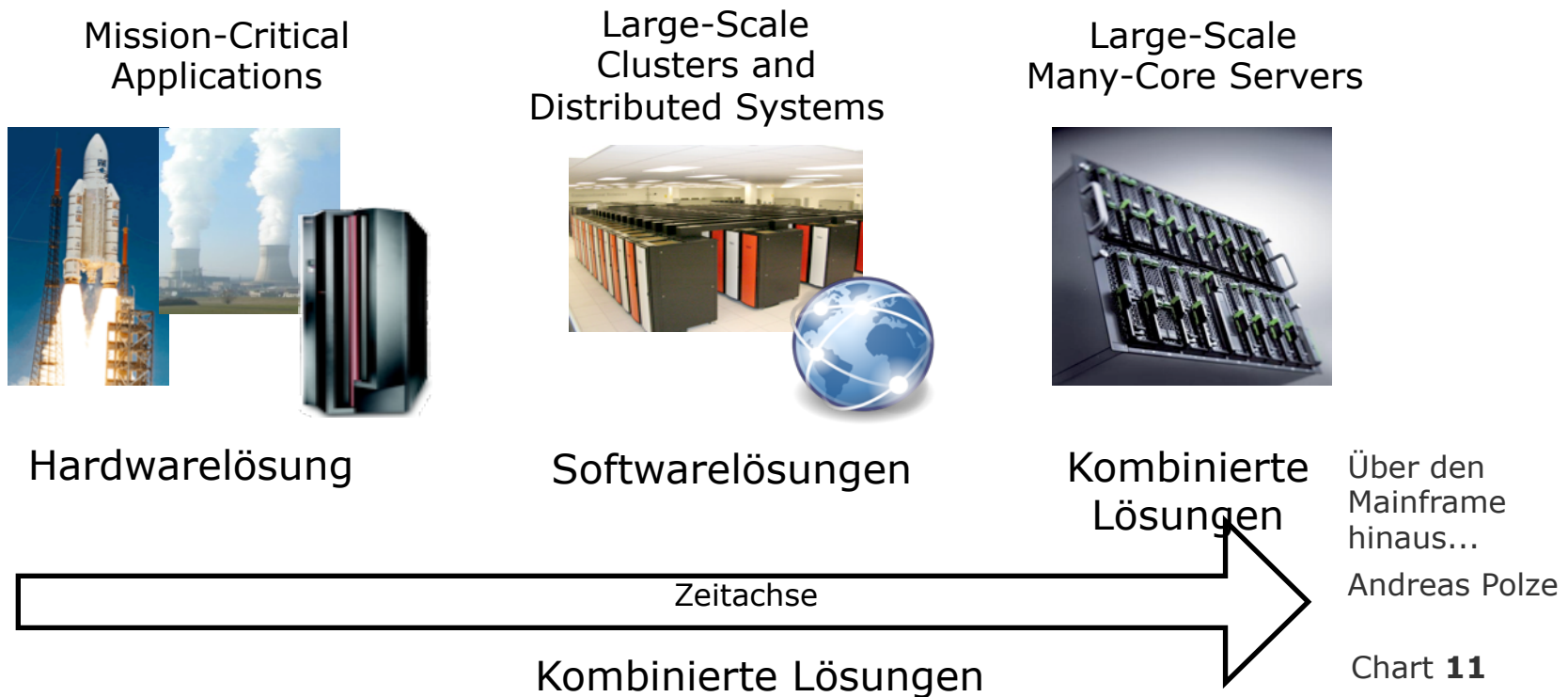


Über den
Mainframe
hinaus...

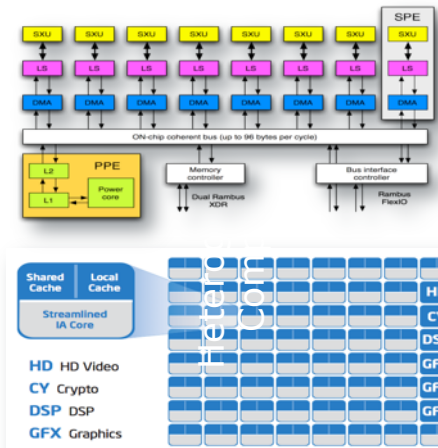
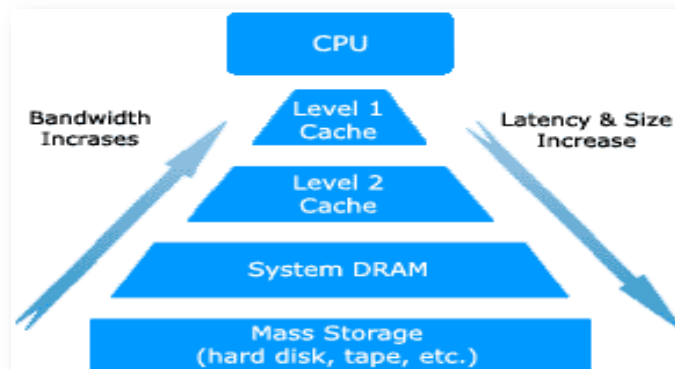
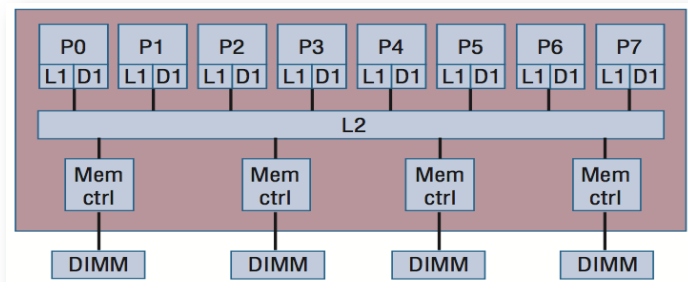
Andreas Polze

Chart **10**

Dependability Research



Revolution bei Standard-Servern (x86/x64)

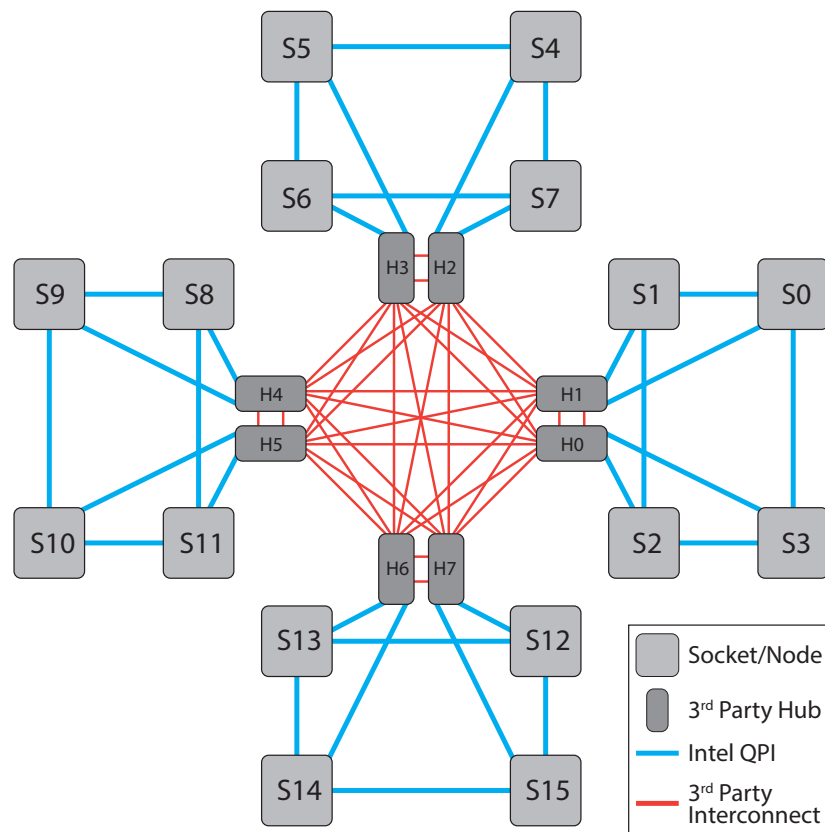


Über den Mainframe hinaus...

Andreas Polze

Chart 12

Core to Cloud: Rackscale NUMA



- SGI UV-300H
- 16 sockets
- 240 cores
- 12 TB memory

Über den
Mainframe
hinaus...

Andreas Polze

Chart **13**

IBM Power E880

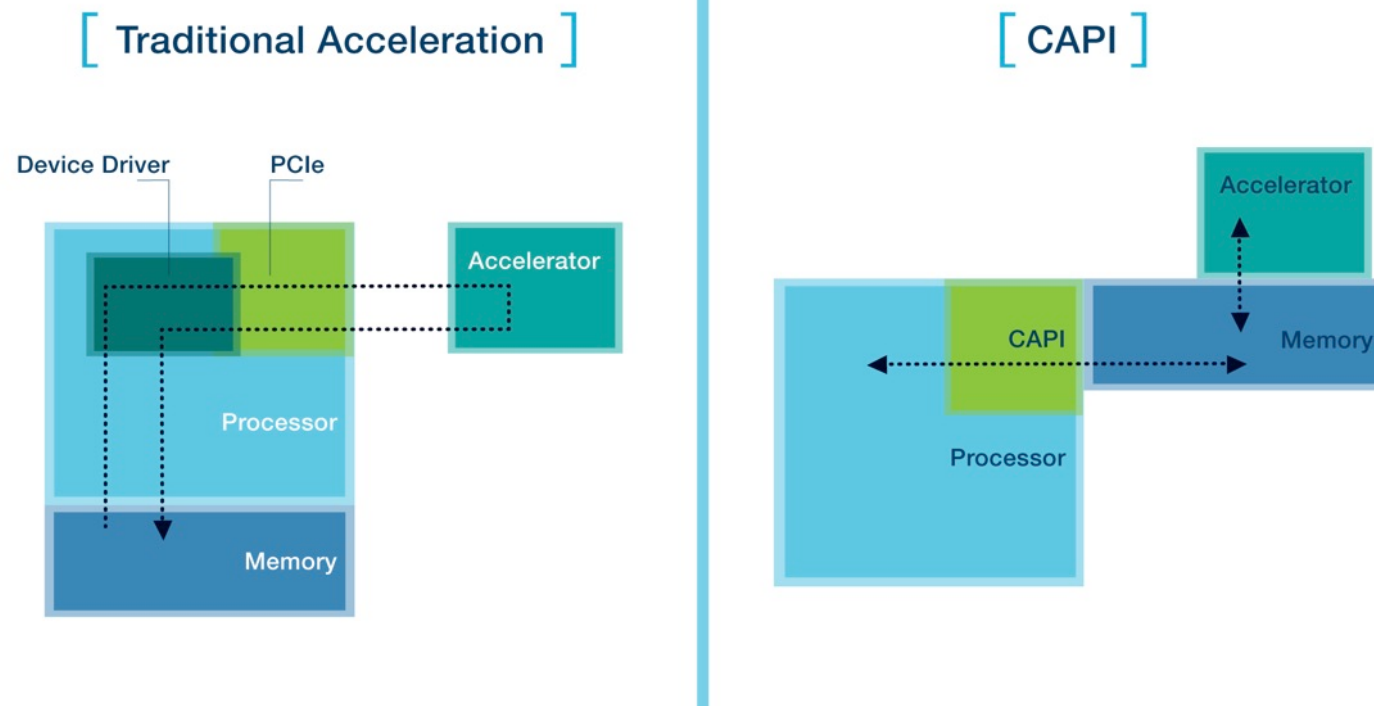


**Über den
Mainframe
hinaus...**

Andreas Polze

Chart **14**

Heterogeneous Computing: The Coherent Accelerator Processor Interface (CAPI)



**Über den
Mainframe
hinaus...**

Andreas Polze

Chart **15**

Reliability: Alte Wahrheiten neu überdacht

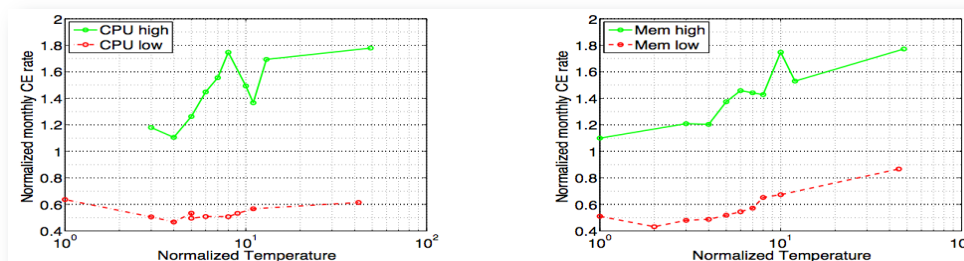
Dramatic shift in single machine reliability aspects

- SMP becomes heterogeneous tiled on-chip network
- Decreasing structural sizes + dynamic frequency and voltage
- Massive memory increase

More fault classes, less error containment !

Few research results from HPC perspective

- Type and intensity of workload significantly influences life time
- Failure rates depend on processor count, not hardware type



Über den
Mainframe
hinaus...

Andreas Polze

Chart **16**

Beobachtungen

Traditional hardware fault models need an update

- Memory with increased density and data rates
- Group of ‚simple‘ cores instead of monolithic processor
- Interconnect as crucial component, fault isolation issues

Reactive fault tolerance gets inappropriate

- Recovery time correlates with system size
- 24/7 business availability demands pro-active fault tolerance
- Reactive does not scale (Example: HPC)

Virtualization as new system layer

- Dependability of (hardware-supported) hypervisors

Weak tool support for reliability research

- Missing consideration of ‚below-OS‘ testing

Über den
Mainframe
hinaus...

Andreas Polze

Chart **17**

FutureSOC Lab

HPI FutureSOC Lab

- Collaboration with industry for software research on next-generation X86 hardware (32-65 cores, 1-2 TB RAM)

Active research work

- Gain understanding of new fault classes
- Failure prediction based on cross-level monitoring data analysis
- Pro-active virtual machine migration
- Fault injection based on UEFI firmware technology
- Parallelized build processing



Über den
Mainframe
hinaus...

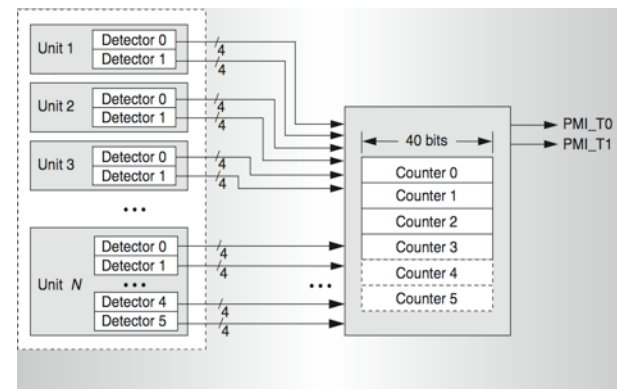
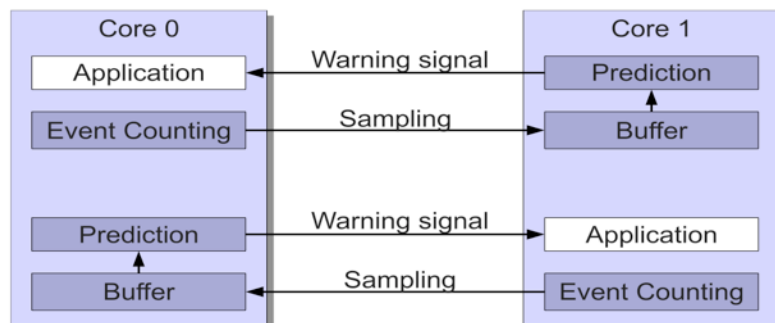
Andreas Polze

Chart **18**

CPU level: Online Hardware Failure Prediction

Using X86 hardware performance events

- Instruction retirement, cache miss, branch miss-prediction, ...
 - Limited number of hardware counter units -> exploit event correlations
 - Threshold-triggered, time-triggered
- Applicable to major cellular multiprocessing platforms (Intel, AMD, SPARC, IBM Power)



Über den
Mainframe
hinaus...

Andreas Polze

Chart 19

Memory level: Beobachtungen im FutureSOC Lab

Date	Severity	Event	Source	Description
15-Jun-2010 13:47:12	Info	No	BIOS	System boot (POST complete)
15-Jun-2010 13:45:53	Major	No	[0x00:00]	POST - 'MEM4_DIMM-2D' memory training failed
15-Jun-2010 13:45:53	Major	No	[0x00:00]	POST - 'MEM4_DIMM-1D' memory training failed
15-Jun-2010 13:45:53	Major	No	[0x00:00]	POST - 'MEM4_DIMM-2B' memory training failed
15-Jun-2010 13:45:53	Major	No	[0x00:00]	POST - 'MEM4_DIMM-1B' memory training failed
15-Jun-2010 13:45:53	Critical	Yes	SMI	'MEM4_DIMM-1D' Memory: Uncorrectable error (ECC)
15-Jun-2010 13:45:53	Critical	Yes	SMI	'MEM4_DIMM-1C' Memory: Uncorrectable error (ECC)
15-Jun-2010 13:45:53	Critical	Yes	SMI	'MEM4_DIMM-1B' Memory: Uncorrectable error (ECC)
15-Jun-2010 13:45:53	Critical	Yes	SMI	'MEM4_DIMM-1A' Memory: Uncorrectable error (ECC)
15-Jun-2010 13:45:40	Critical	Yes	iRMC S2	'MEM4_DIMM-2D': Memory module failed (disabled)
15-Jun-2010 13:45:40	Critical	Yes	iRMC S2	'MEM4_DIMM-1D': Memory module failed (disabled)
15-Jun-2010 13:45:40	Critical	Yes	iRMC S2	'MEM4_DIMM-2B': Memory module failed (disabled)
15-Jun-2010 13:45:40	Critical	Yes	iRMC S2	'MEM4_DIMM-1B': Memory module failed (disabled)
15-Jun-2010 13:43:43	Info	No	BIOS	System boot (POST complete)
14-Jun-2010 17:41:47	Critical	Yes	iRMC S2	'MEM4_DIMM-1D': Memory module error
14-Jun-2010 17:26:17	Major	Yes	iRMC S2	'MEM4_DIMM-1D': Memory module failure predicted

Über den
Mainframe
hinaus...

Andreas
Polze

Chart 20

z/PDT + z/OS R11: the mainframe in our lab

The IBM System z Personal Development Tool (zPDT)

An overview

The zPDT technology consists of hardware and software

- **zPDT software enables System z architecture**
- **zPDT hardware (1090-L01, -L02, -L03, USB hardware key) is a security key that authenticates the zPDT software**
 - Plugs into the USB port (looks like a 'memory stick')
 - The key expires annually and must be re-certified annually
 - The zPDT software will not run without the 1090 USB key installed
 - The 1090 can enable an environment with 1, 2, or 3 virtual engines



A working zPDT-based application development system consists of many pieces:

- Thinkpad or System x (or equivalent) ... and ...
- 64-bit Red Hat or openSUSE Linux on the Thinkpad or System x ... and ...
- zPDT hardware (1090 – in USB port) ... and ...
- zPDT software ... and ...
- System z operating system(s) / middleware / tools



OS level:

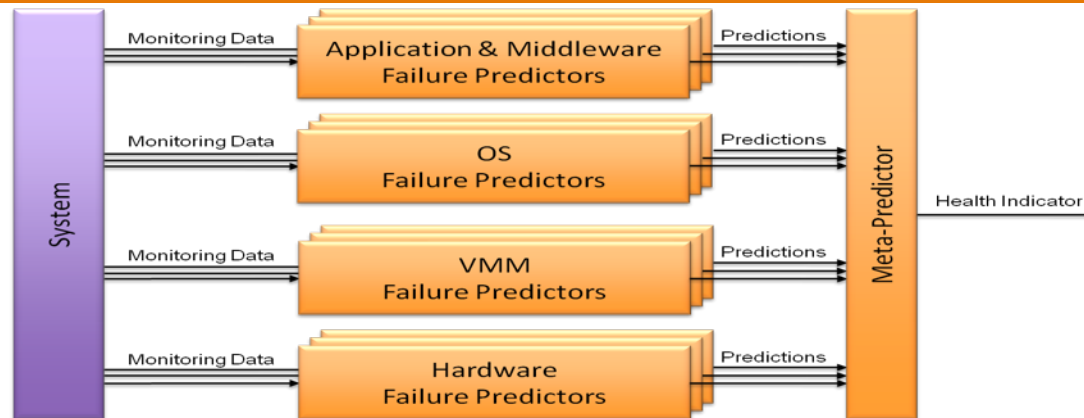
Predictive Failure Analysis (PFA) and
Runtime Diagnostics (RD) in z/OS R11

Über den
Mainframe
hinaus...

Andreas Polze

Chart **21**
(sick-but-not-death-incidents)

Fehlervorhersage: Maschinelles Lernen



Ensemble learning:

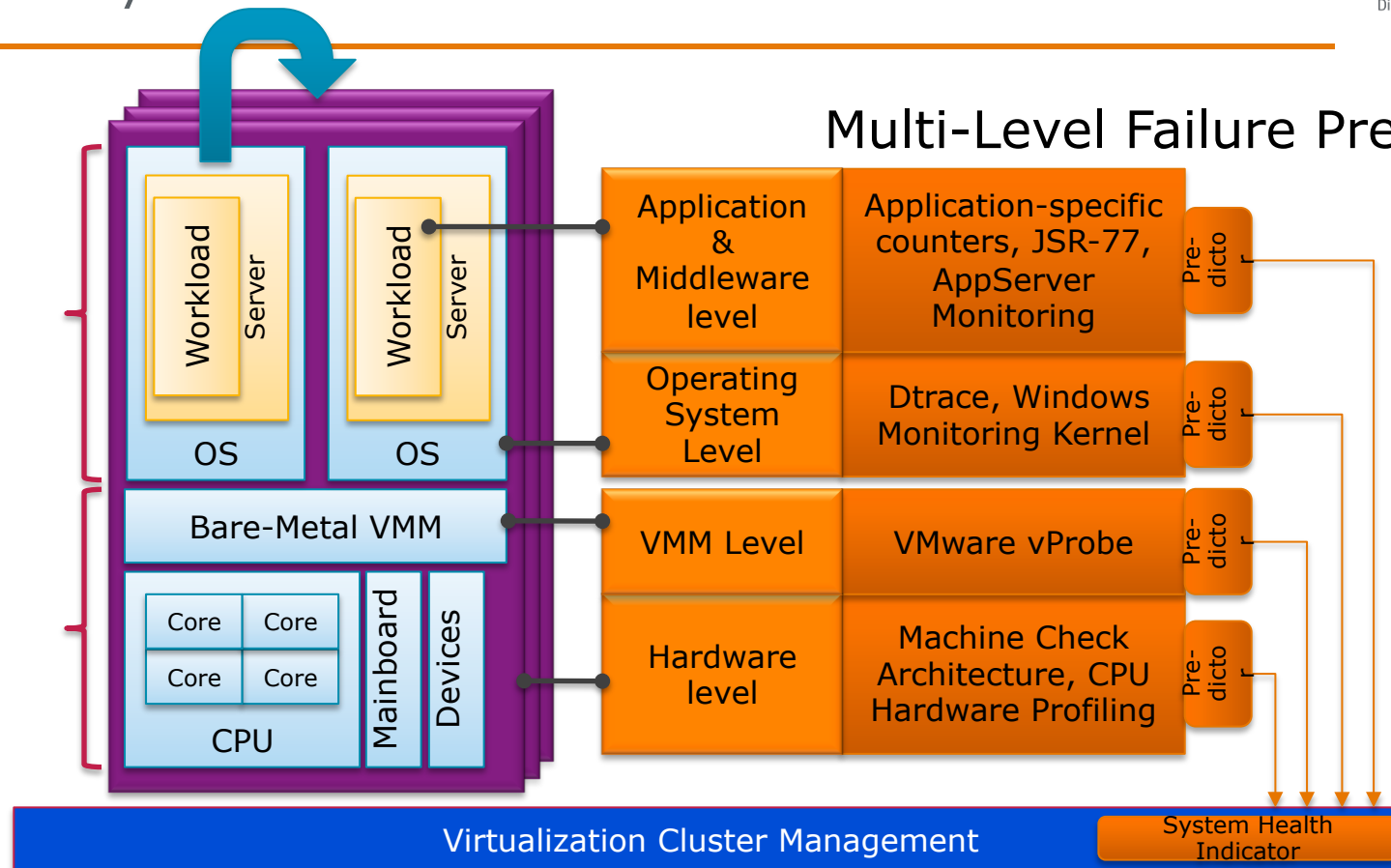
- Boosts accuracy – which failure-prone situations can best be identified by either hardware, OS, VMM failure predictors?
- Domain knowledge – operating system vendors know their system best and can provide the most advanced predictor on OS level
- Pluggable – domain predictors provided by an application vendor can easily be integrated into our anticipatory virtualization architecture
- Ensemble-learning can combine predictions across all system levels

Über den
Mainframe
hinaus...

Andreas Polze

Chart **22**

Global System Health Indicator

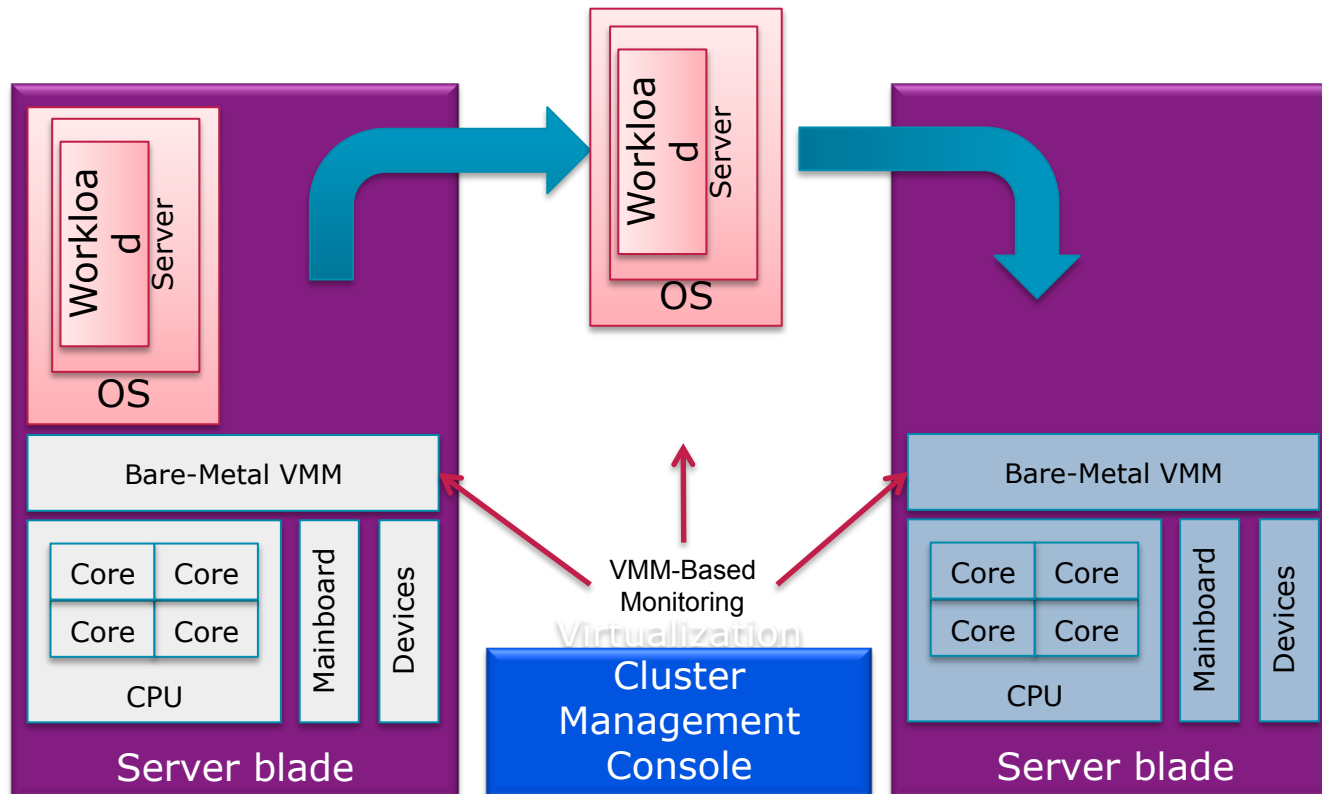


Über den Mainframe hinaus...
Andreas Polze

Chart **23**

VM Migration aus Basis von Fehlervorhersage

Power HA, Parallel Sysplex



Über den
Mainframe
hinaus...

Andreas Polze

Chart **24**

Schlußfolgerungen

- Server entwickeln sich...
 - Immer größere Zahl von CPU cores
 - Riesige Mengen an Speicher (DRAM)
 - Nichtflüchtiger Speicher (NVRAM, SSD)
- Zuverlässigkeit wird zum Schlüsselkriterium
 - Hohe Integration und sinkende Strukturgrößen (22nm, 14nm) zukünftiger CPUs führen zu Mehr-Bit-Fehlern
 - Fehlervorhersage und dynamische Rekonfiguration (dLPAR, VM Migration) als Lösungsansätze
- Isolation gegen Fehlerausbreitung (LPARs, blades)

Zukünftige Server benötigen neue Programmier- und Managementmodelle

- Microservice-Architectures

Über den
Mainframe
hinaus...

Andreas Polze

Chart **25**

Computer architecture drives changes in system software

Servers have evolved...

- New form factors
- Higher density
- Standard architectures
- Multicore/multithreaded

Advances in operating systems

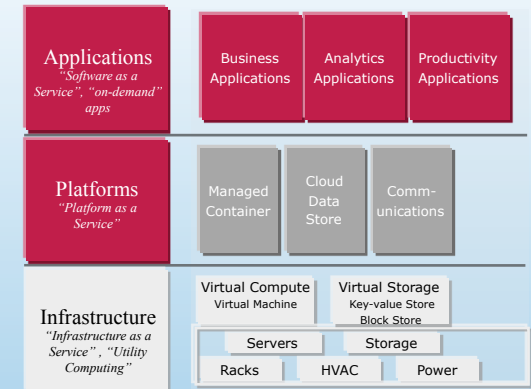
- Virtualization
- Thrustworthiness/security
- Clustering
- Need for new programming models, SW Architectures, Services



Cloud Computing – the three layers

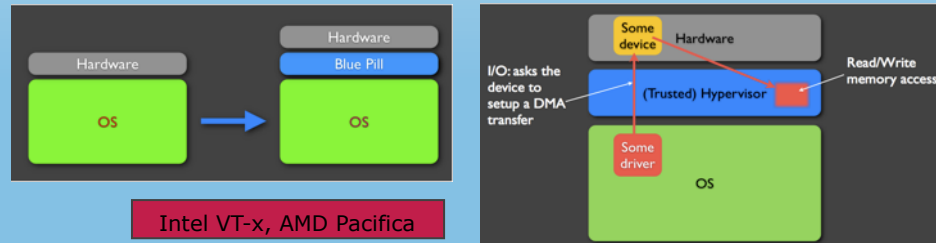
Challenges:

- Has to abstract underlying hardware
- Be elastic in scaling to demand
- Pay per use basis

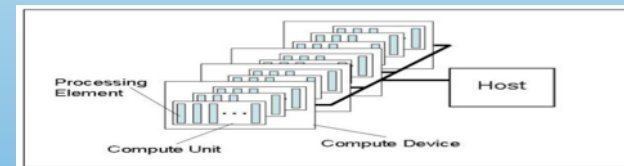


Virtualization problems

- Security: extended attack surface
- Virtualization-based malware
- Must trust hypervisor



Hybrid Computing OpenCL: New Programming Models



One Host + one or more Compute Devices

- Each Compute Device is composed of one or more Compute Units
- Each Compute Unit is further divided into one or more Processing Elements



Die Mainframe setzt den Gold-Standard im Rechenzentrum

openHPI.de