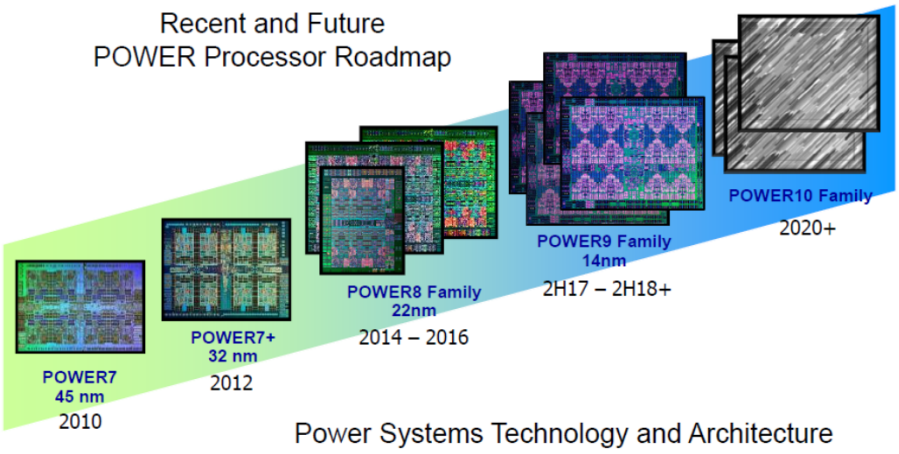
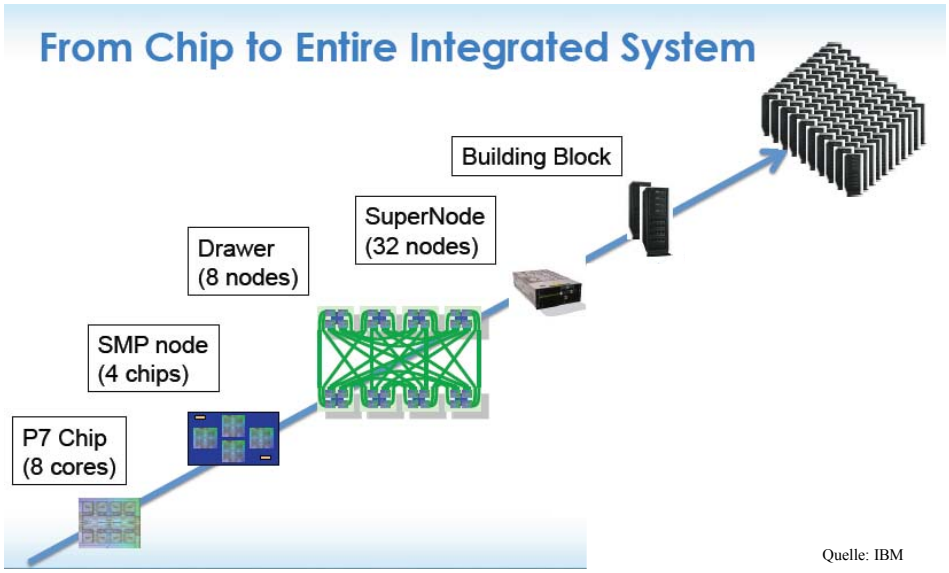
Example: IBM POWER Processor



Source: IBM

IBM Power Supercomputer



Quelle: IBM

Power9 Cores with Simultaneos Multithreading

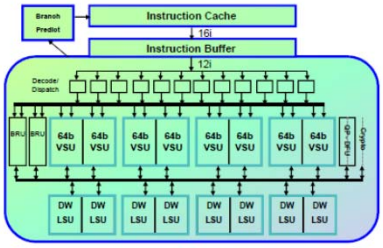
Optimized for Stronger Thread Performance and Efficiency

- Increased execution bandwidth efficiency for a range of workloads including commercial, cognitive and analytics
- Sophisticated instruction scheduling and branch prediction for unoptimized applications and interpretive languages
- Adaptive features for improved efficiency and performance especially in lower memory bandwidth systems

Available with SMT8 or SMT4 Cores

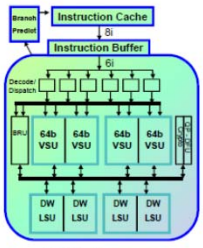
8 or 4 threaded core built from modular execution slices

POWER9 SMT8 Core
- PowerVM Ecosystem Continuity
- Strongest Thread
- Optimized for Large Partitions



SMT8 Core

POWER9 SMT4 Core
- Linux Ecosystem Focus
- Core Count / Socket
- Virtualization Granularity

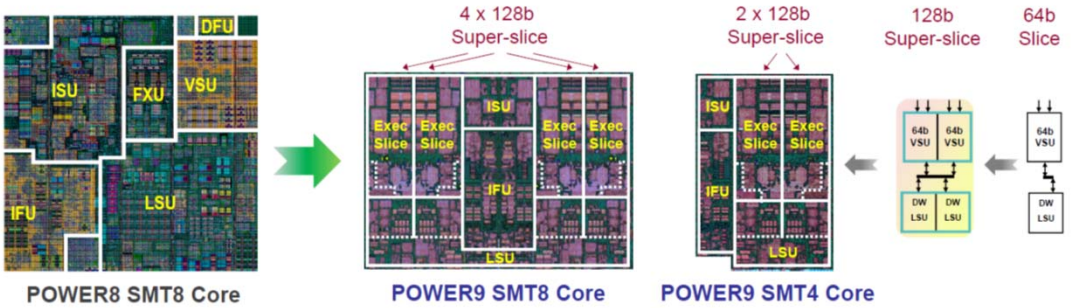


SMT4 Core

Source: IBM

POWER9 Core Microarchitecture

Modular Execution Slices



POWER8 SMT8 Core | POWER9 SMT8 Core | POWER9 SMT4 Core

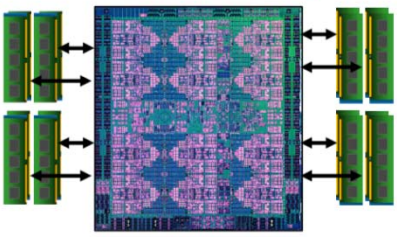
Re-factored Core Provides Improved Efficiency & Workload Alignment

- Enhanced pipeline efficiency with modular execution and intelligent pipeline control
- Increased pipeline utilization with symmetric data-type engines: Fixed, Float, 128b, SIMD
- Shared compute resource optimizes data-type interchange

Source: IBM

POWER9 – Dual Memory Subsystem

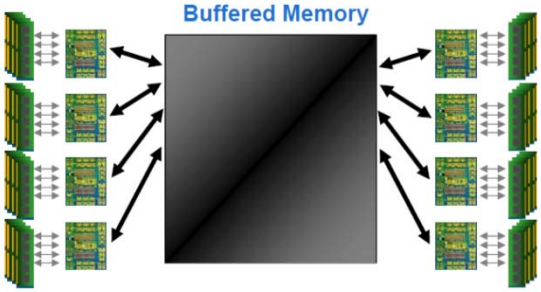
Scale Out
Direct Attach Memory



8 Direct DDR4 Ports

- Up to 120 GB/s of sustained bandwidth
- Low latency access
- Commodity packaging form factor
- Adaptive 64B / 128B reads

Scale Up
Buffered Memory



8 Buffered Channels

- Up to 230GB/s of sustained bandwidth
- Extreme capacity – up to 8TB / socket
- Superior RAS with chip kill and lane sparing
- Compatible with POWER8 system memory
- Agnostic interface for alternate memory innovations

Source: IBM

POWER9 Processor – Common Features

New Core Microarchitecture

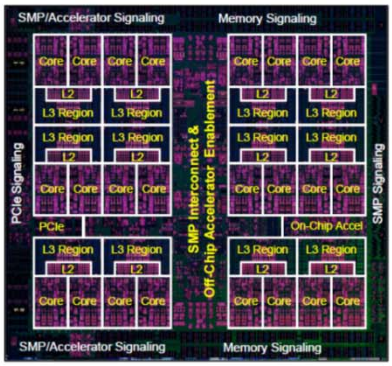
- Stronger thread performance
- Efficient agile pipeline
- POWER ISA v3.0

Enhanced Cache Hierarchy

- 120MB NUCA L3 architecture
- 12 x 20-way associative regions
- Advanced replacement policies
- Fed by 7 TB/s on-chip bandwidth

Cloud + Virtualization Innovation

- Quality of service assists
- New interrupt architecture
- Workload optimized frequency
- Hardware enforced trusted execution



14nm finFET Semiconductor Process

- Improved device performance and reduced energy
- 17 layer metal stack and eDRAM
- 8.0 billion transistors

Leadership Hardware Acceleration Platform

- Enhanced on-chip acceleration
- Nvidia NVLink 2.0: High bandwidth and advanced new features (25G)
- CAPI 2.0: Coherent accelerator and storage attach (PCIe G4)
- New CAPI: Improved latency and bandwidth, open interface (25G)

State of the Art I/O Subsystem

- PCIe Gen4 – 48 lanes

High Bandwidth Signaling Technology

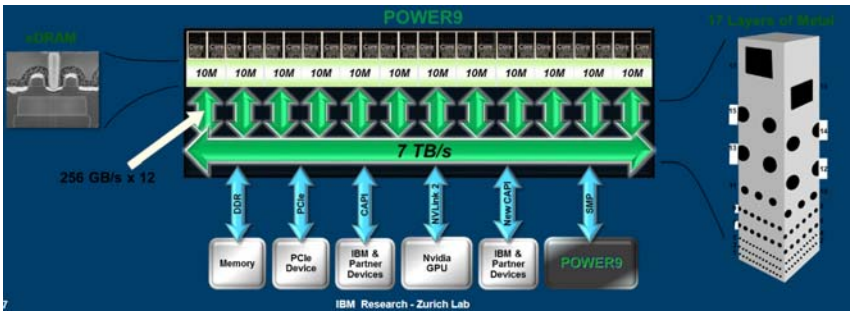
- 16 Gb/s interface
  - Local SMP
- 25 Gb/s Common Link interface
  - Accelerator, remote SMP

Source: IBM

POWER9 Data Capacity & Throughput

L3 Cache

- 120 MB shared capacity
- Per 2x Core
  - 10 MB L3 cache region
  - 512 kB L2 cache



High-throughput on-chip fabric

- Over 7 TB/s on-chip switch
- Move data in/out at 256 GB/s per 2x Core

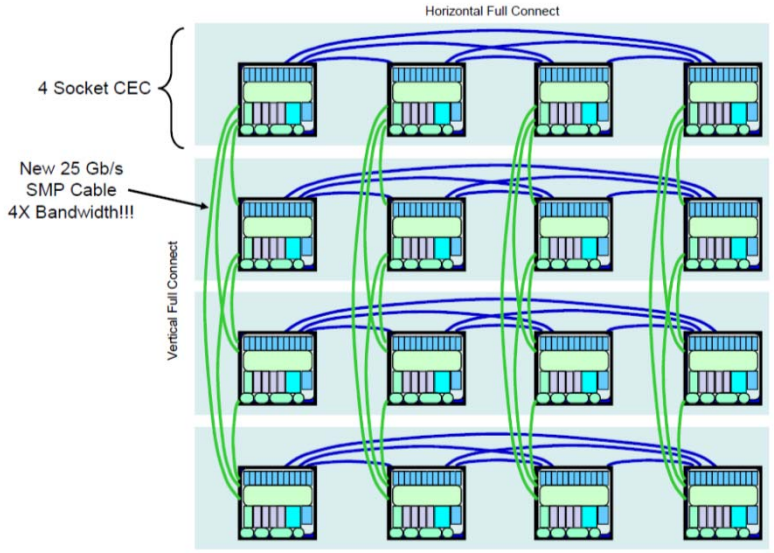
POWER9 – Scale Out Family

L922 9008-22L	S922 9009-22A	S914 9009-41A	S924 9009-42A	H922 9223-22H	H924 9223-42H
• 1,2-socket, 2U • 8,10,12 cores/skt • 32 IS DIMM slots • 4TB memory • 4 CAPI 2.0 Slots • Linux only • PowerVM • KVM (GA2)	• 1,2-socket, 2U • 4, 8,10 cores/skt • 32 IS DIMM slots • 4TB memory • 4 CAPI 2.0 Slots • AIX, IBM i, & Linux • PowerVM	• 1-socket, 4U & Tower • 4,6,8 cores/skt • 16 IS DIMM slots • 1TB memory • 2 CAPI 2.0 Slots • Internal RDX Media • AIX, IBM i, Linux • PowerVM	• 2-socket, 4U • 8,10,12 cores/skt • 32 IS DIMM slots • 4TB memory • 4 CAPI 2.0 slots • Internal RDX Media • AIX, IBM i, Linux • PowerVM	• 1,2-socket, 2U • 4, 8,10 cores/skt • 32 IS DIMM slots • 4TB memory • 4 CAPI 2.0 Slots • AIX, IBM i up to 25% • Linux • PowerVM	• 2-socket, 4U • 8,10,12 cores/skt • 32 IS DIMM slots • 4TB memory • 4 CAPI 2.0 slots • Internal RDX Media • AIX, IBM i up to 25% • Linux • PowerVM
FAMILY FEATURES	• **Cloud enabled - Embedded virtualization capabilities with PowerVM** • **DDR4 Industry Standard (IS) memory RDIMMs** • **High Speed 25Gb/s external ports – one per socket** • **2 Internal NVMe Flash boot adapters** • **Embedded Analytics and Algorithms on the chip help run POWER9 at an always optimized frequency** • **No internal DVD Drive**				

Source: IBM

Paderborn Center for Parallel Computing

POWER9 – 16 Socket 2-Hop System Topology



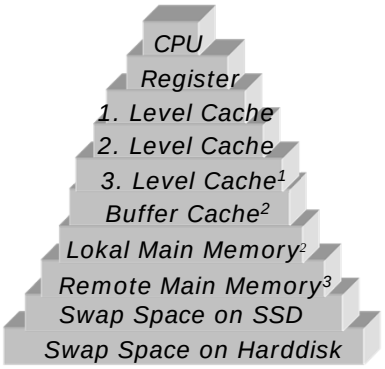
Source: IBM

Interconnection Network

- HUB/Switch (one per SMP node)
  - 192 GB/s to host node
  - 336 GB/s to 7 other nodes in same drawer
  - 240 GB/s to 24 nodes in other 3 drawers in same SuperNode
  - 320 GB/s to hubs in other SuperNodes

Source: IBM

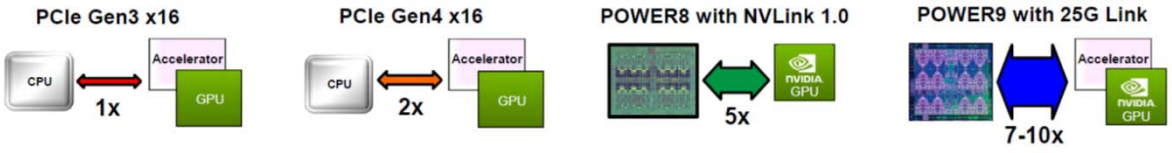
Memory Hierarchy: Example IBM Power E870 (Power8)



Kapazität	Bandbreite	Latenz
256 Byte	120 GByte/s	0.2 ns
64 kByte	75 GByte/s	1 ns
512 kByte	150 GByte/s	4 ns
80 MByte	150 GByte/s	< 30 ns
128 Mbyte	?	?
1024 GByte	230 GByte/s	< 90 ns
8192 GByte	230 GByte/s	< 1 µs
>	X * 500 MByte/s	< 1 ms
>>	X * 200 MByte/s	~5 ms

[1] 8MB per Core x 10
[2] shared by 10 Cores
[2] shared by 80 Cores

POWER9 - Acceleration



Increased Performance / Features / Acceleration Opportunity

Extreme Accelerator Bandwidth and Reducend Latencys

- PCIe Gen 4 x 48 lanes – 192 GB/s peak bandwidth
- IBM BlueLink 25 Gb/s x 48 lanes – 300 GB/s peak bandwidth

Coherent Memory and Virtual Addressing Capability for all Accelerators

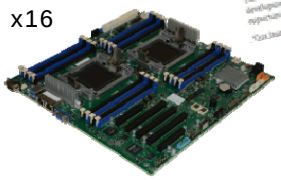
- CAPI 2.0 using PCIe Gen 4
- NVLink 2.0 next generation of GPU/CPU bandwidth and integration usinmg BLueLink
- OpenCAPI – openinterface with high bandwidth and low latency using BlueLink

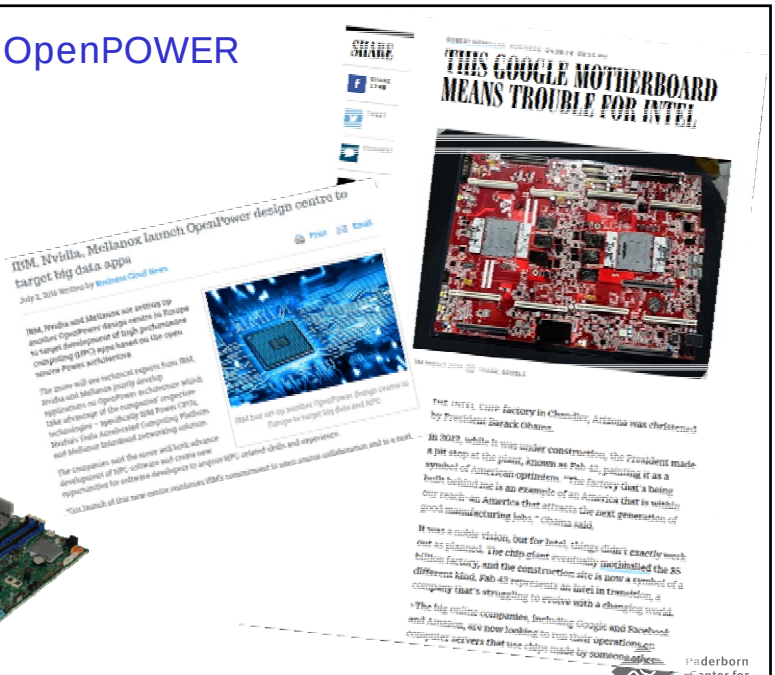
Source: IBM

OpenPOWER

Compared to IBM products
- Broader market
- Bigger ecosystem
- Platform for innovation
- Main focus on Linux

- Raptor Computing Systems
  - Talos-II
  - Two Power9 processors
  - 16 DIMMs ECC DDR4
  - 3 x PCIe 4.0 x16

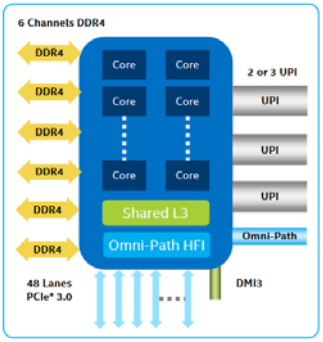




Example: Intel Xeon "Skylake"

- AVX-512 – 64 Single-Precision FLOP/s or 32 Double-Precision FLOP/s
- Ultra Path Interconnect (UPI) with 10.4 Gigatransfers per second (GT/s)

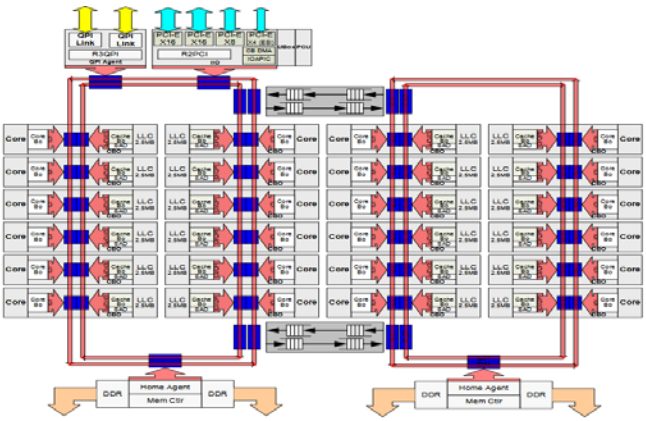
Features	Intel® Xeon® Processor E5-2600 v4	Intel® Xeon® Scalable Processor
Cores Per Socket	Up to 22	Up to 28
Threads Per Socket	Up to 44 threads	Up to 56 threads
Last-level Cache (LLC)	Up to 55 MB	Up to 38.5 MB (non-inclusive)
QPI/UPI Speed (GT/s)	2x QPI channels @ 9.6 GT/s	Up to 3x UPI @ 10.4 GT/s
PCIe* Lanes/ Controllers/Speed(GT/s)	40 / 10 / PCIe* 3.0 (2.5, 5, 8 GT/s)	48 / 12 / PCIe 3.0 (2.5, 5, 8 GT/s)
Memory Population	4 channels of up to 3 RDIMMs, LRDIMMs, or 3DS LRDIMMs	6 channels of up to 2 RDIMMs, LRDIMMs, or 3DS LRDIMMs
Max Memory Speed	Up to 2400	Up to 2666
TDP (W)	55W-145W	70W-205W



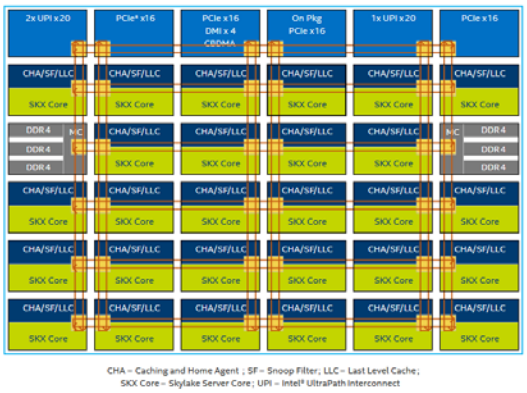
Source: Intel

Intel Xeon SP – Mesh Interconnect Architecture

Mesh improves scalability with higher bandwidth and reduced latencies



Broadwell EX 24-core



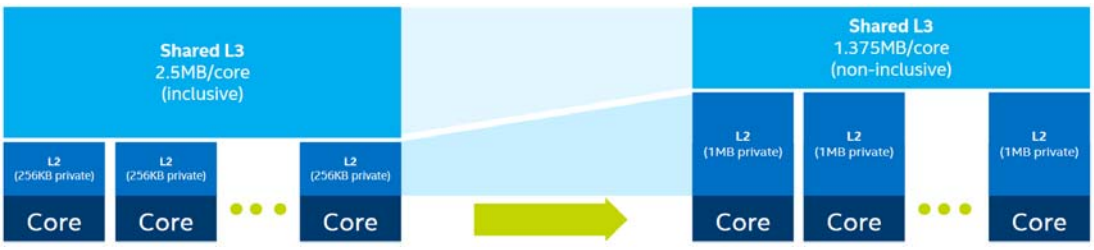
Skylake SP 28-core

Source: Intel

Intel Xeon SP – Cache Hierarchy

On-chip cache

- Processor core with
  - 640 KiB L1 data cache and 640 KiB L1 instruction cache (both 8-way set associative)
  - 1 MiB L2 cache (16-way set associative)
- private L2 becomes primary cache with shared L3 used as overflow cache
- Non-inclusive L3 cache (1.375 MiB / core) – lines in L2 may not exist in L3



previous Broadwell architecture

Skylake SP architecture

Source: Intel

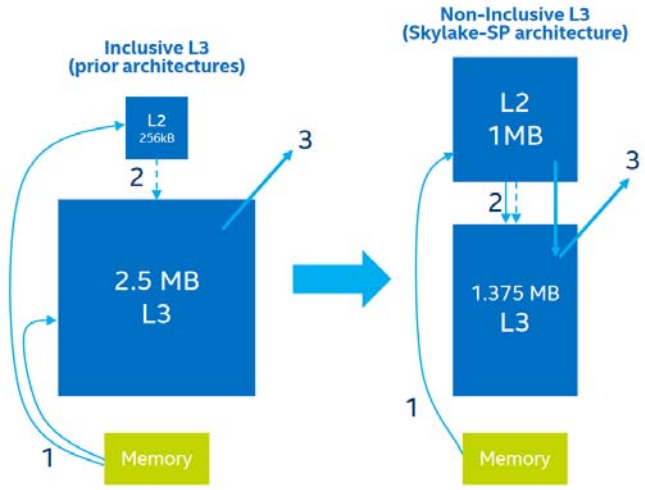
Intel Xeon SP Family

Xeon E5 to Xeon Scalable Normalized Comparison Chart

E5 v3	CC | TDP | GHz	SIR*	SFR*	LINPACK*		E5 v4	CC | TDP | GHz	SIR*	SFR*	LINPACK*		Skylake-SP	CC | TDP | GHz	SIR*	SFR*	LINPACK*	
E5-2699 v3	18C | 145W | 2.3	51.6%	54.6%	36.9%	→	E5-2699A v4	22C | 145W | 2.4	67.6%	65.9%	43.9%	→	8180	28C | 205W | 2.5	100.0%	100.0%	100.0%	Platinum
						E5-2699 v4	22C | 145W | 2.2	63.2%	64.6%	41.6%	→	8176	28C | 165W | 2.1	89.7%	92.3%	76.5%	
E5-2698 v3	16C | 135W | 2.3	48.0%	52.4%	32.9%	→	E5-2698 v4	20C | 135W | 2.2	58.4%	63.4%	37.8%	→	8170	26C | 165W | 2.1	84.5%	89.0%	71.0%	
												8168	24C | 205W | 2.7	92.4%	95.7%	95.8%	
						E5-2697A v4	16C | 145W | 2.6	57.2%	61.2%	37.0%	→	8164	26C | 150W | 2	82.2%	88.6%	65.5%	
E5-2697 v3	14C | 145W | 2.6	47.6%	52.1%	33.9%	→	E5-2697 v4	18C | 145W | 2.3	59.2%	62.7%	37.8%	→	8160	24C | 150W | 2.1	79.8%	87.5%	70.6%	
												6154	18C | 200W | 3	77.2%	85.5%	79.4%	
E5-2695 v3	14C | 120W | 2.3	43.2%	49.1%	29.4%	→	E5-2695 v4	18C | 120W | 2.1	53.6%	59.0%	32.1%	→	6152	22C | 140W | 2.1	74.2%	83.3%	64.7%	
E5-2690 v3	12C | 135W | 2.6	42.4%	48.4%	29.6%	→	E5-2690 v4	14C | 135W | 2.6	51.6%	57.3%	30.9%	→	6150	18C | 165W | 2.7	72.8%	82.7%	71.8%	
E5-2687W v3	10C | 160W | 3.1	38.2%	46.5%	0.0%	→	E5-2687W v4	12C | 160W | 3.2	46.8%	54.7%	32.8%	→	6148	20C | 150W | 2.4	73.9%	83.3%	67.2%	
												6148	20C | 150W | 2.4	73.9%	83.3%	67.2%	
												6146	12C | 165W | xx				
												6142	16C | 150W | 2.6	64.7%	75.6%	53.8%	
E5-2683 v3	14C | 120W | 2	42.4%	49.0%	29.9%	→	E5-2683 v4	16C | 120W | 2.1	48.4%	55.7%	28.6%	→	6138	20C | 125W | 2	66.3%	76.9%	54.6%	Gold
E5-2680 v3	12C | 120W | 2.5	40.4%	46.9%	27.6%	→	E5-2680 v4	14C | 120W | 2.4	46.8%	54.3%	27.9%	→	6140	18C | 140W | 2.3	66.4%	77.4%	56.7%	
E5-2670 v3	12C | 120W | 2.3	38.4%	46.4%	27.4%	→							6136	12C | 150W | 3	56.3%	68.3%	52.9%	
												6144	8C | 165W | xx				
												6132	14C | 133W | 2.6	58.5%	69.6%	50.0%	
E5-2667 v3	8C | 135W | 3.2	31.9%	39.6%	0.0%	→	E5-2667 v4	8C | 135W | 3.2	35.6%	44.7%	21.8%	→	6130	16C | 125W | 2.1	57.6%	70.4%	43.7%	
E5-2660 v3	10C | 105W | 2.6	34.9%	43.0%	24.5%	→	E5-2660 v4	14C | 105W | 2	43.6%	53.0%	25.0%	→	6130	16C | 125W | 2.1	57.6%	70.4%	43.7%	
E5-2650 v3	10C | 105W | 2.3	32.7%	41.9%	24.2%	→	E5-2650 v4	12C | 105W | 2.2	39.3%	49.4%	22.7%	→	6126	12C | 125W | 2.6	50.6%	62.8%	42.9%	
E5-2650L v3	12C | 65W | 1.8	29.0%	36.2%	0.0%	→	E5-2650L v4	14C | 65W | 1.7	37.0%	45.7%	17.6%	→	6134	8C | 130W | 3.3	41.2%	53.6%	35.3%	
E5-2643 v3	6C | 135W | 3.4	26.4%	34.9%	0.0%	→	E5-2643 v4	6C | 135W | 3.4	27.9%	37.5%	17.6%	→	5120T	14C | 105W | xx				
E5-2640 v3	8C | 90W | 2.6	27.8%	35.6%	19.5%	→	E5-2640 v4	10C | 90W | 2.4	33.1%	42.7%	21.0%	→	5118	12C | 105W | xx				
E5-2637 v3	4C | 135W | 3.5	18.2%	25.8%	0.0%	→	E5-2637 v4	4C | 135W | 3.5	18.5%	26.8%	13.4%	→	5119T	14C | 70W | xx				
												6128	6C | 115W | xx				
E5-2630L v3	8C | 55W | 1.8	20.6%	27.9%	0.0%	→	E5-2630L v4	10C | 55W | 1.8	24.6%	33.5%	13.7%	→	4116	12C | 85W | xx				
E5-2630 v3	8C | 85W | 2.4	26.4%	33.5%	0.0%	→	E5-2630 v4	10C | 85W | 2.2	31.3%	40.1%	18.9%	→	5122	4C | 105W | 3.6	22.4%	31.5%	22.7%	
E5-2623 v3	4C | 105W | 3	16.6%	23.8%	0.0%	→	E5-2623 v4	4C | 85W | 2.6	15.0%	22.1%	9.2%	→	41xx	8C | 85W | xx				
E5-2620 v3	6C | 85W | 2.4	20.1%	27.4%	0.0%	→	E5-2620 v4	8C | 85W | 2.1	25.2%	35.1%	15.1%	→	4114T	10C | 70W | xx				Silver
												4114	10C | 85W | xx				
												4112	4C | 85W | xx				
												4108	8C | 85W | xx				
												41xx	8C | 85W | xx				
E5-2609 v3	6C | 85W | 1.9	12.2%	19.9%	0.0%	→	E5-2609 v4	8C | 85W | 1.7	14.8%	23.1%	14.3%	→	3106	8C | 85W | xx				Bronze
E5-2603 v3	6C | 85W | 1.6	10.6%	18.2%	0.0%	→	E5-2603 v4	6C | 85W | 1.7	11.4%	19.1%	10.7%	→	3104	6C | 85W | xx				

Source: Intel

Inclusive vs Non-Inlcusive L3



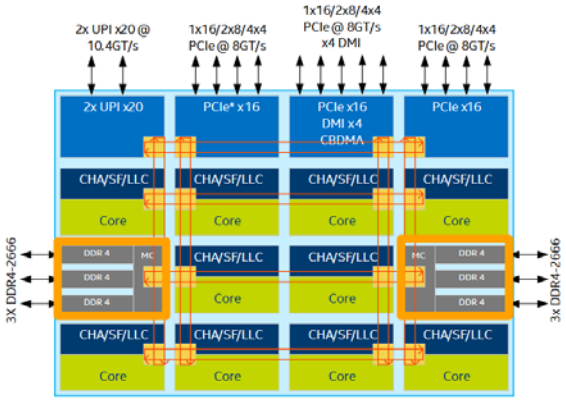
Source: Intel

1) memory reads fill directly to the L2, no longer to both the L2 and L3
2) When a L2 line needs to be removed, both modified and unmodified lines are written back
3) Data shared across cores are copied into the L3 for servicing future L2 misses

Closer look in a later lecture

Source: Intel
Source: Intel

Memory Subsystem



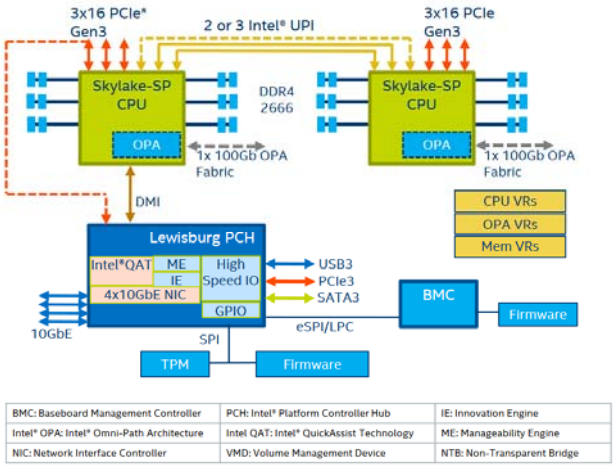
Source: Intel

2 Memory Controllers, 3 channels each

- DDR4 up to 2666, 2 DIMMs per channel
- 1.5 TB max memory capacity per socket

Source: Intel
Source: Intel

Intel Xeon Scalable Processor

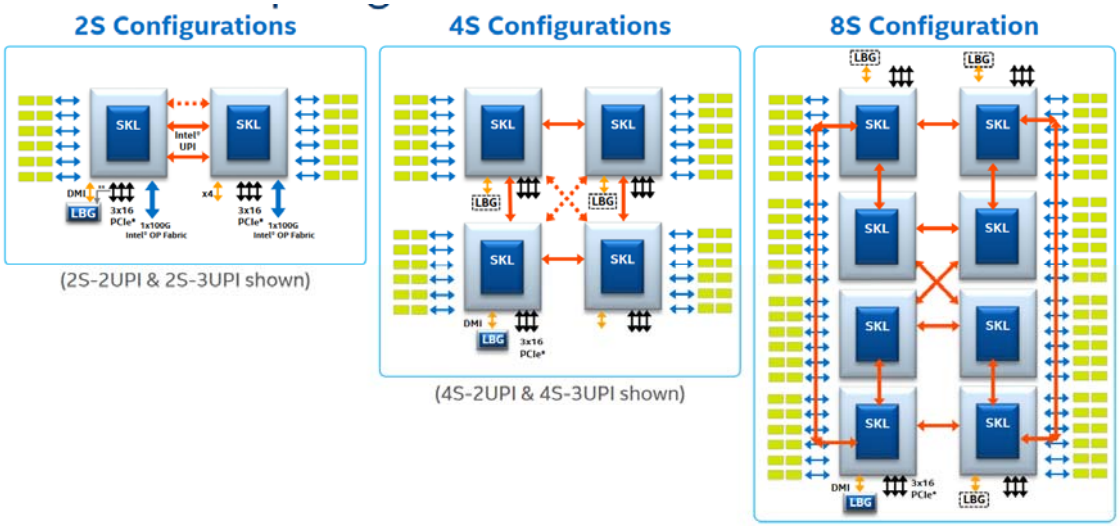


BMC: Baseboard Management Controller	PCH: Intel® Platform Controller Hub	IE: Innovation Engine
Intel® OPA: Intel® Omni-Path Architecture	Intel QAT: Intel® QuickAssist Technology	ME: Manageability Engine
NIC: Network Interface Controller	VMD: Volume Management Device	NTB: Non-Transparent Bridge

Feature	Details
Socket	Socket P
Scalability	2S, 4S, 8S, and >8S (with node controller support)
CPU TDP	70W – 205W
Chipset	Intel® C620 Series (code name Lewisburg)
Networking	Intel® Omni-Path Fabric (integrated or discrete) 4x10GbE (integrated w/ chipset) 100G/40G/25G discrete options
Compression and Crypto Acceleration	Intel® QuickAssist Technology to support 100Gb/s comp/decomp/crypto 100K RSA2K public key
Storage	Integrated QuickData Technology, VMD, and NTB Intel® Optane™ SSD, Intel® 3D-NAND NVMe & SATA SSD
Security	CPU enhancements (MBE, PPK, MPX) Manageability Engine Intel® Platform Trust Technology Intel® Key Protection Technology
Manageability	Innovation Engine (IE) Intel® Node Manager Intel® Datacenter Manager

Source: Intel

Intel Xeon SP - Platform Topologies

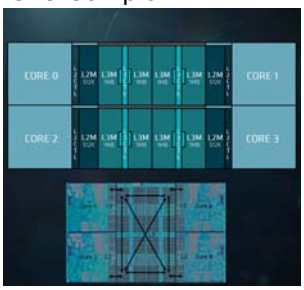


Source: Intel

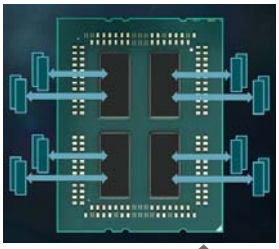
Example: AMD Epyc 7000 Series

- ZEN Microarchitecture
  - L1 D-cache with 32 kiB, 8 way
  - L1 I-cache with 64 kiB, 4 way
  - L2 cache with 512 kiB, 8 way
- CPU Complex
  - Four cores connected to an L3 cache
  - L3 cache with 8 MiB, 16 way associative
- Multi chip processors
  - Four CCX per processor
- Infinity Fabric
  - 42 GiB/s bi-directional bandwidth per link
  - Fully connected coherent Infinity Fabric within socket
  - Dual socket systems with two processors connected with 4 x 38 GiB/s links

CPU Complex



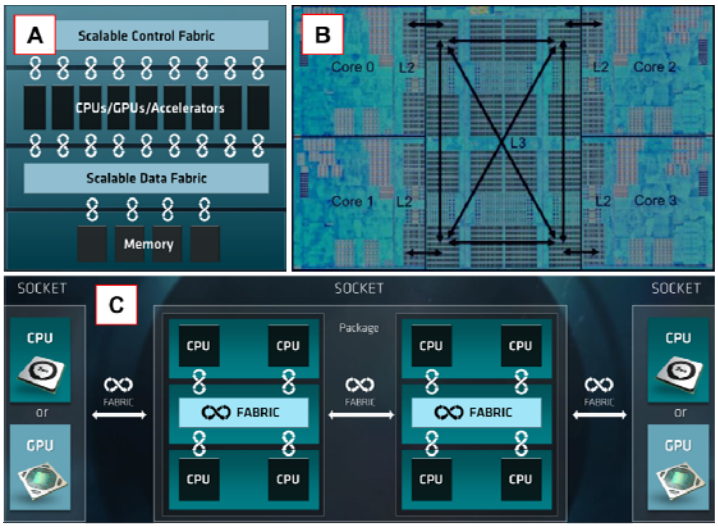
Multi Chip processor

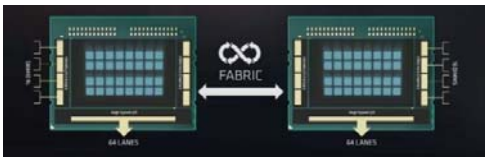


AMD Epyc 7000 Series

- AMD EPYC 7601
  - 32 Cores, 2.2 GHz (max boost clock 3.2 GHz, all cores max boost 2.7 GHz)
  - 64 MiB L3-cache
  - TDP 180 Watt
  - 1 or 2 sockets
- AMD EPYC 7451
  - 24 cores, 2.3 GHz (max boost clock 3.2 GHz)
  - TDP 180 Watt
  - 1 or 2 sockets

AMD Infinity Fabric





Two socket platform

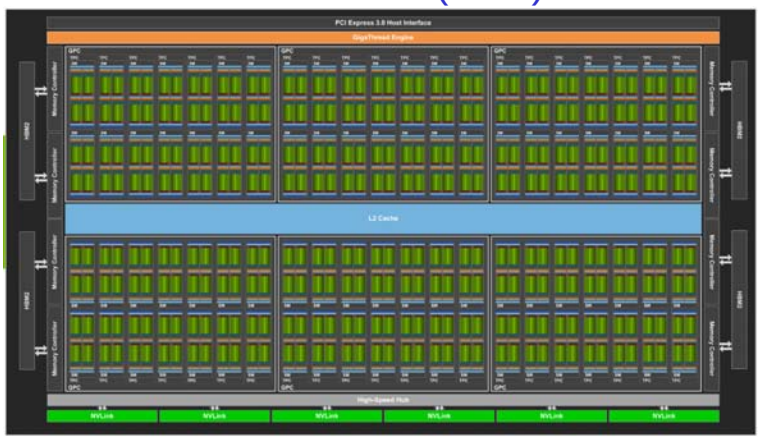
AMD Infinity Fabric connecting Zeppelin die on a MCM and between MCMs

Source: AMD

Example: NVIDIA Stream Processor (GPU)

NVIDIA Tesla V100

- 21 billion transistors
- 80 SM stream multi-processors
  - 5,120 CUDA Cores
  - 1.45 GHz
  - 6 MiB shared L2 cache
- 640 tensor cores
  - Accelerates Deep learning applications
- Main memory
  - 16 GiB HBM2 (High-Bandwidth-Memory)
  - 900 GiB/s
- NVLINK

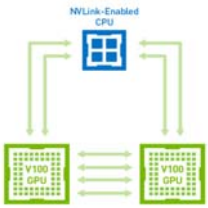


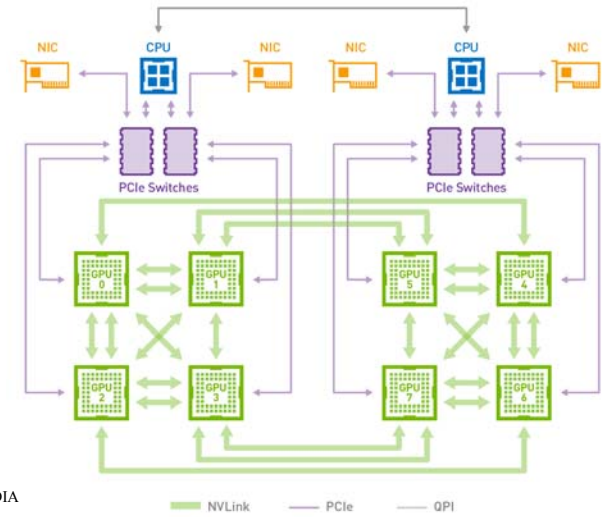
Source: NVIDIA

- 7.5 TFLOPS DP, 15 TFLOPS SP
- Training, Inference: 120 TOPS

NVIDIA NVLINK

- 6 NVLINKs per GPU
- 300 GiB/s bandwidth



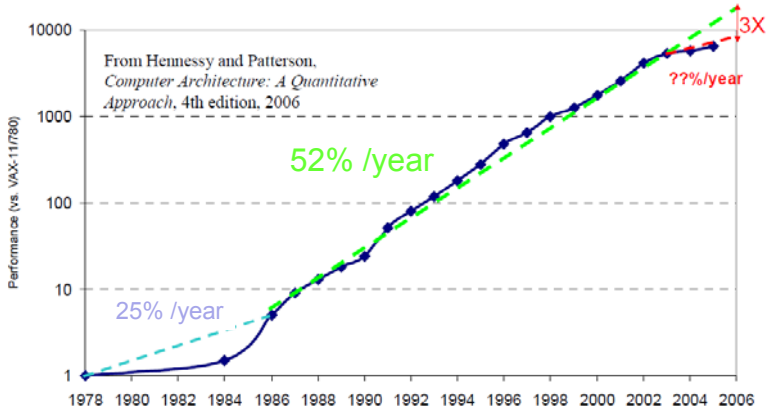


Source: NVIDIA

Accelerators become part of the Processor

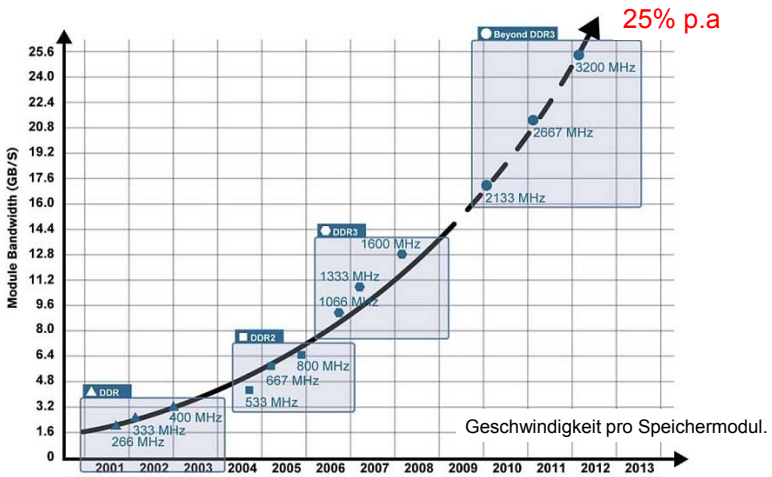
- Floating-Point Unit
  - 1978: Intel 8086 + Intel 8087 Math-Co processor (16 Bit)
  - 1989: Intel i486 with integrated floating-point units (32 bit)
- Vector Unit
  - 1993: CM5 with Sparc processor + Vector Unit Accelerators (MBUS)
  - 1995: Intel Pentium P55C with MMX instructions
  - 1996: Motorola PowerPC with AltiVec
- Stream Processing
  - 2006: Workstation + GPU graphic card (PCI)
  - 2011: Intel HD Graphics 3000 with integrated GPU (OpenCL)

Leistungsentwicklung eines Prozessorkerns



- Von 1986 bis 2002 ca. 50% Leistungszuwachs pro Jahr
- Derzeit Einzelprozessorleistung nur langsam zunehmend
- Höherer Leistungszuwachs nur noch über Erhöhung der Anzahl an Prozessoren (Cores) möglich

Hauptspeichergeschwindigkeit



Leistungsunterschied zwischen CPU und RAM wird weiter wachsen (52% p.a vs 25% p.a)

Quelle: Rambus Inc., 2010

Memory Bandwidth/Latency

Generation	Type	Peak Bandwidth	Latency (1st word)
SDRAM (1990s)	PC-100	0.8 Gbyte/s	20 ns
DDR (2000)	DDR-200	1.6 Gbyte/s	20 ns
DDR	DDR-400	3.2 Gbyte/s	15 ns
DDR2 (2003)	DDR2-667	5.3 Gbyte/s	15 ns
DDR2	DDR2-800	6.4 Gbyte/s	15 ns
DDR3 (2007)	DDR3-1066	8.5 Gbyte/s	13 ns
DDR3	DDR3-1600	12.8 Gbyte/s	11.25 ns
DDR4 (2014)	DDR4-2133	17 Gbyte/s	~ 11 ns
DDR4	DDR4-2666	21 Gbyte/s	~10.5 ns

Trends

- „Power Wall“
  - Energieaufnahme / Kühlung
  - Lösungen
    - geringere Taktfrequenzen
    - mehr Ausführungseinheiten
- „Memory Wall“
  - Speicherbandbreite u. Latenz
  - Lösungen
    - bessere Speicherhierarchien u. Anbindung an CPUs
    - Latency-Hidding
- „ILP Wall“
  - Beschränkte Parallelität im sequentiellen Instruktionsstrom
  - Lösungen
    - mehr Parallelität in Programmen erkennen (Compiler)
    - mehr explizite Parallelität in Programmen (Programmiersprachen)

Architekturen paralleler Rechnersysteme

Einfache Definition Parallelrechner

George S. Almasi, *IBM Thomas J. Watson Research Center*
Allan Gottlieb, *New York University, 1989*

„ *A parallel computer is a collection of processing elements that communicate and cooperate to solve large problems fast.*"

Rechnerarchitektur allgemein

Eine Rechnerarchitektur ist bestimmt durch ein Operationsprinzip für die Hardware und die Struktur ihres Aufbaus aus den einzelnen Hardware-Betriebsmitteln

(Giloi 1993)

Operationsprinzip

Das Operationsprinzip definiert das funktionelle Verhalten der Architektur durch Festlegung einer Informationsstruktur und einer Kontrollstruktur.

Hardware-Struktur

Die Struktur einer Rechnerarchitektur ist gegeben durch Art und Anzahl der Hardware-Betriebsmittel und deren verbindenden Kommunikationseinrichtungen.

… in anderen Worten

Operationsprinzip

- Vorschrift über das Zusammenspiel der Komponenten

Struktur

- Einzelkomponenten und deren Verknüpfung

- Grundlegende Strukturbausteine sind
  - Prozessor (CPU), als aktive Komponente zur Ausführung von Programmen,
  - Hauptspeicher (ggf. hierarchisch strukturiert, …),
  - Übertragungsmedium zur Verbindung der einzelnen Architekturkomponenten,
  - Steuereinheiten für Anschluss und Kontrolle von Peripherie-geräten und
  - Geräte, als Zusatzkomponenten für Ein- und Ausgabe von Daten sowie Datenspeicherung.

Parallelrechner

- Operationsprinzip:
  - gleichzeitige Ausführung von Befehlen
  - sequentielle Verarbeitung in bestimmbaren Bereichen

- Arten des Parallelismus:
  - Explizit: Die Möglichkeit der Parallelverarbeitung wird a priori festgelegt. Hierzu sind geeignete Datentypen bzw. Datenstrukturen erforderlich, z.B. Vektoren (lineare Felder) samt Vektoroperationen.
  - Implizit: Die Möglichkeit der Parallelverarbeitung ist nicht a priori bekannt. Durch eine Datenabhängigkeitsanalyse werden die parallelen und sequentiellen Teilschritte des Algorithmus zur Laufzeit ermittelt.

Strukturelemente von Parallelrechnern

- Parallelrechner besteht aus einer Menge von Verarbeitungselementen, die in einer koordinierten Weise, teilweise zeitgleich, zusammenarbeiten, um eine Aufgabe zu lösen

- Verarbeitungselemente können sein:
  - spezialisierte Einheiten, wie z.B. die Pipeline-Stufen eines Skalarprozessors oder die Vektor-Pipelines der Vektoreinheit eines Vektorrechners
  - gleichartige Rechenwerke, wie z.B. die Verarbeitungselemente eines Feldrechners
  - Prozessorknoten eines Multiprozessorsystems
  - vollständige Rechner, wie z.B. Workstations oder PCs eines Clusters
  - selbst wieder ganze Parallelrechner oder Cluster

Grenzbereiche von Parallelrechnern

- eingebettete Systeme als spezialisierte Parallelrechner
- Superskalar-Prozessoren, die feinkörnige Parallelität durch Befehls-Pipelining und Superskalar-Technik nutzen
- Mikroprozessoren arbeiten als Hauptprozessor teilweise gleichzeitig zu einer Vielzahl von spezialisierten Einheiten wie der Bussteuerung, DMA-,Graphikeinheit, usw.
- Ein-Chip-Multiprozessor
- mehrfädige (multithreaded) Prozessoren führen mehrere Kontrollfäden überlappt oder simultan innerhalb eines Prozessors aus
- VLIW- (Very Long Instruction Word)- Prozessor

Klassifikation von Parallelrechnern

- Klassifikation nach Flynn, d.h. Klassifikation nach der Art der Befehlsausführung
- Klassifikation nach der Speicherorganisation und dem Adressraum
- Konfigurationen des Verbindungsnetzwerks
- Varianten an speichergekoppelte Multiprozessorsysteme
- Varianten an nachrichtengekoppelte Multiprozessorsysteme

Klassifikation nach Flynn

Zweidimensionale Klassifizierung mit Kriterium Anzahl der Befehls- und Datenströme

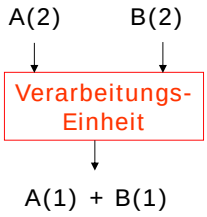
- Rechner bearbeitet zu einem Zeitpunkt einen oder mehrere Befehle
- Rechner bearbeitet zu einem Zeitpunkt einen oder mehrere Datenwerte

⇒ Damit vier Klassen von Rechnerarchitekturen

- **SISD**: Single Instruction, Single Data
  Ein Befehl verarbeitet einen Datensatz. (herkömmliche Rechnerarchitektur eines seriellen Rechners)
- **SIMD**: Single Instruction, Multiple Data
  Ein Befehl verarbeitet mehrere Datensätze, z.B. N Prozessoren führen zu einem Zeitpunkt den gleichen Befehl aber mit unterschiedlichen Daten aus.
- **MISD**: Multiple Instruction, Single Data
  Mehrere Befehle verarbeiten den gleichen Datensatz. (Diese Rechnerarchitektur ist nie realisiert worden.)
- **MIMD**: Multiple Instruction, Multiple Data
  Unterschiedliche Befehle verarbeiten unterschiedliche Datensätze.
  (Dies ist das Konzept fast aller modernen Parallelrechner.)

SISD Architektur

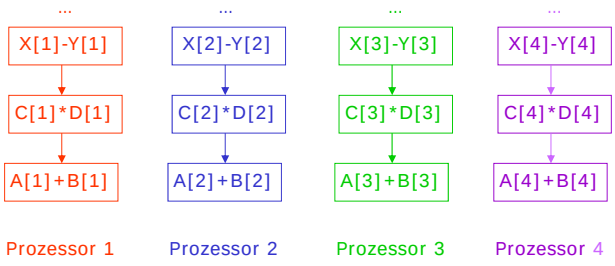
- Klassische Struktur eines seriellen Rechners:
  Nacheinander werden verschiedene Befehle ausgeführt, die z.B. einzelne Datenpaare verknüpfen



- Moderne RISC (Reduced Instruction Set Computer) Prozessoren verwenden Pipelining:
  - Mehrere Funktionseinheiten, die gleichzeitig aktiv sind.
  - Operationen sind in Teiloperationen unterteilt.
  - In jedem Takt kann eine Funktionseinheit (z.B. Addititionseinheit) eine neue Operation beginnen.
  - D.h. hohe interne Parallelität nutzbar

SIMD Architektur (Prozessorarray)

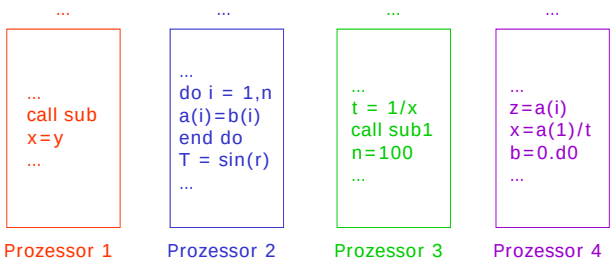
- Mehrere Prozessoren führen zu einem Zeitpunkt den gleichen Befehl aus
- Rechner für Spezialanwendungen (z.B. Bildverarbeitung, Spracherkennung)
- I.A. sehr viele Prozessorkerne (tausende Kerne in einem System)
- Beispiele: Graphikprozessoren, Numerische Coprozessoren



- Mittlerweile auch innerhalb einzelner Funktionseinheiten zu finden

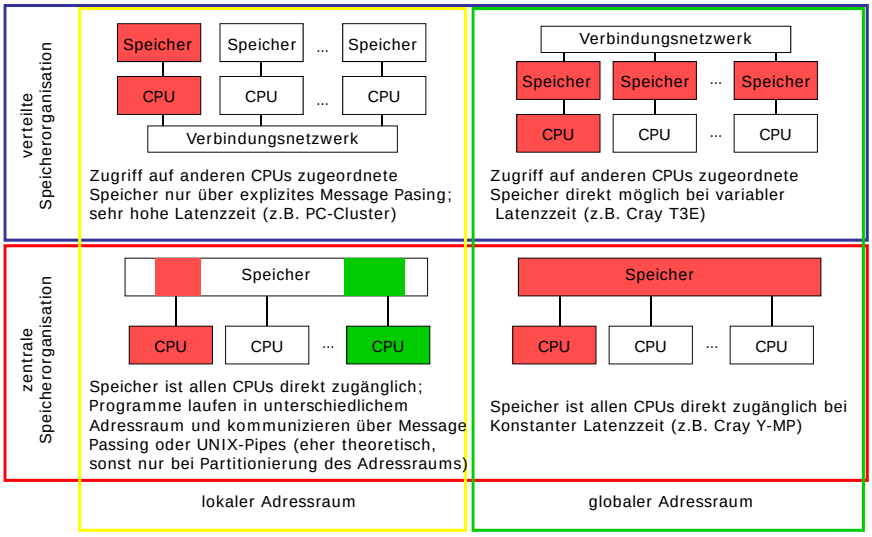
MIMD Architektur

Mehrere Prozessoren führen unabhängig voneinander unterschiedliche Instruktionen auf unterschiedlichen Daten aus:

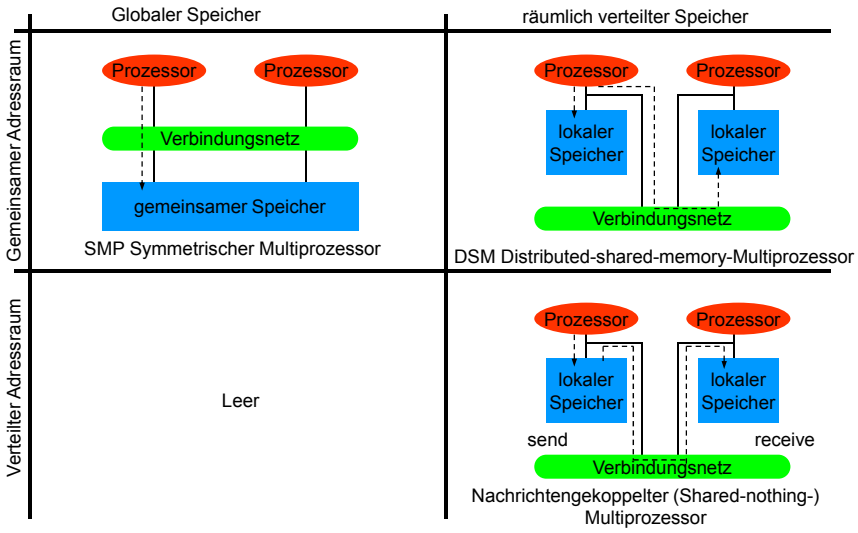


- Fast alle aktuellen Systeme entsprechen dieser Architektur.

Speicherorganisation und Adressraum



Konfiguration der Verbindungsnetzwerke



Arten von Multiprozessorsystemen

- Bei speichergekoppelten Multiprozessorsystemen besitzen alle Prozessoren einen gemeinsamen Adressraum.
  Kommunikation und Synchronisation geschehen über gemeinsame Variablen.
  - **symmetrisches Multiprozessorsystem (SMP)**: ein globaler Speicher
  - **Distributed-Shared-Memory-System (DSM)**: gemeinsamer Adressraum trotz räumlich verteilter Speichermodule

- Beim nachrichtengekoppelten Multiprozessorsystem besitzen alle Prozessoren nur räumlich verteilte Speicher und prozessorlokale Adressräume.
  Die Kommunikation geschieht durch Austausch von Nachrichten.
  - Massively Parallel Processors (MPP), eng gekoppelte Prozessoren
  - Verteiltes Rechnen in einem **Workstation-Cluster**.
  - **Grid-/Cloud-Computing**: Zusammenschluss weit entfernter Rechner

Speichergekoppelte Multiprozessorsysteme

- Alle Prozessoren besitzen **einen** gemeinsamen Adressraum; Kommunikation und Synchronisation geschieht über **gemeinsame Variablen**.
- Uniform-Memory-Access-Modell (UMA):
  - Alle Prozessoren greifen in gleichermaßen auf einen gemeinsamen Speicher zu. Insbesondere ist die *Zugriffszeit* aller Prozessoren auf den gemeinsamen Speicher *gleich*.
    Jeder Prozessor kann zusätzlich einen *lokalen Cache-Speicher* besitzen.
    Typische Beispiel: die symmetrischen Multiprossorsysteme (SMP)
- Nonuniform-Memory-Access-Modell (NUMA):
  - Die *Zugriffszeiten* auf Speicherzellen des gemeinsamen Speichers *variieren* je nach dem Ort, an dem sich die Speicherzelle befindet.
    Die Speichermodule des gemeinsamen Speichers sind physisch auf die Prozessoren aufgeteilt.
  - Typische Beispiele: Distributed-Shared-Memory-Systeme.

Nachrichtengekoppelte Multiprozessorsysteme

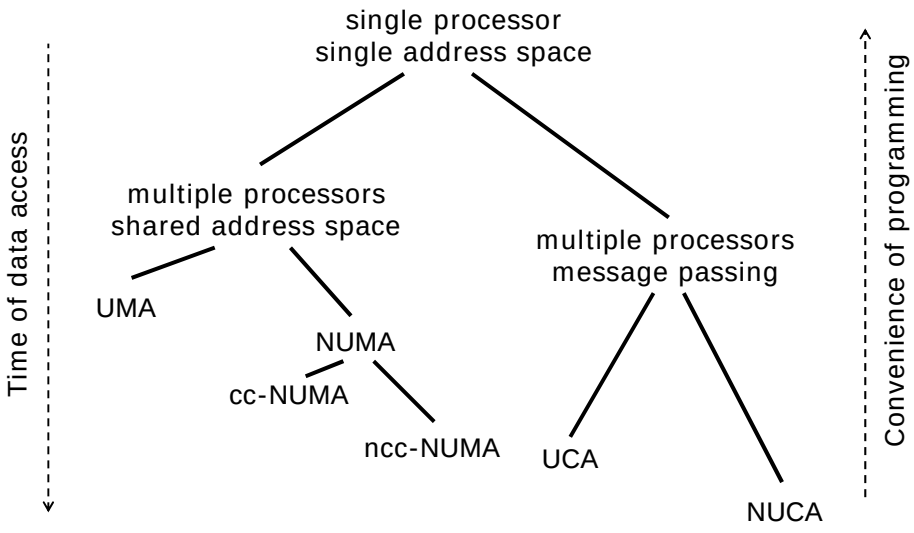
- **Uniform-Communication-Architecture-Modell** (UCA):
  Zwischen allen Prozessoren können gleich lange Nachrichten mit einheitlicher Übertragungszeit geschickt werden.

- **Non-Uniform-Communication-Architecture-Modell** (NUCA):
  Die Übertragungszeit des Nachrichtentransfers zwischen den Prozessoren ist je nach Sender- und Empfänger-Prozessor unterschiedlich lang.

Speicher- vs. Nachrichtenkopplung

- Distributed-Shared-Memory-Systeme sind NUMAs: Die Zugriffszeiten auf Speicherzellen des gemeinsamen Speichers variieren je nach Ort, an dem sich die Speicherzelle befindet.
  - **cc-NUMA** (Cache-coherent NUMA): Cache-Kohärenz wird über das gesamte System gewährleistet, z.B. SGI Origin, HP Superdome, IBM Regatta
  - **ncc-NUMA** (Non-Cache-coherent NUMA): Cache-Kohärenz wird nur innerhalb eines Knoten gewährleistet, z.B. Cray T3E, SCI-Cluster
  - **COMA** (Cache-only-Memory-Architecture): Der Speicher des gesamten Rechners besteht nur aus Cache-Speicher. Nur in einem kommerziellen System realisiert (ehemalige Firma KSR)
- Nachrichten gekoppelte Multiprozessorsysteme sind **NORMAs** (No-remote-memory-access-Modell) oder Shared-nothing-Systeme, z.B. IBM SP, HP Alpha Cluster

Zugriffszeit-/Übertragungszeit-Modell



Zusammenfassung: Klassifizierung

Klassifizierung nach

- Befehls- und Datenströme,
- Speicherorganisation,
- Verbindungsnetzwerk
  - weitere Details später in der Vorlesung

Quantitative Bewertung von Parallelrechnern

Merkmale: Geschwindigkeit, Auslastung

- Ausführungszeit T eines parallelen Programms
  - Zeit zwischen dem Starten der Programmausführung auf einem der Prozessoren bis zu dem Zeitpunkt, an dem der **letzte** Prozessor die Arbeit an dem Programm beendet hat
- Während der Programmausführung sind alle Prozessoren in einem von drei Zuständen
  - rechnend
  - kommunizierend
  - untätig

Ausführungszeit T

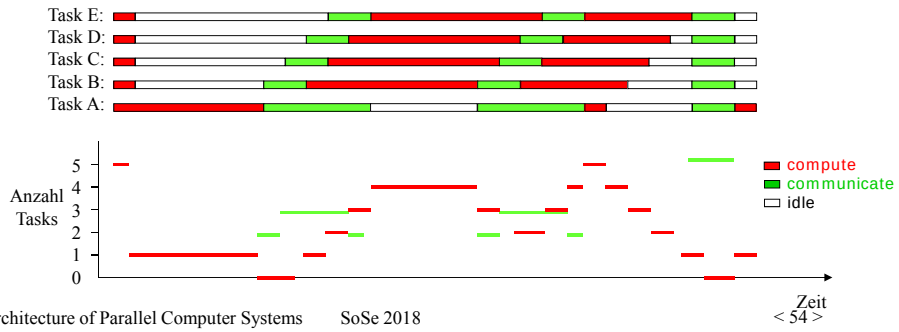
Ausführungszeit T eines parallelen Programms auf einem dediziert zugeordneten Parallelrechner setzt sich zusammen aus:

- Berechnungszeit T_{comp}
  - Zeit für die Ausführung von Rechenoperationen
- Kommunikationszeit T_{com}
  - Zeit für Sende- und Empfangsoperationen
- Untätigkeitszeit T_{idle}
  - Zeit für Warten (auf zu empfangende oder zu sendende Nachrichten)

Es gilt: $T \approx T_{comp} + T_{com} + T_{idle}$

Parallelitätsprofil

- Parallelitätsprofil zeigt die vorhandene Parallelität in einem parallelen Programm (einer konkreten Ausführung)
  - Grafische Darstellung:
    Auf der x-Achse wird die Zeit und auf der y-Achse die Anzahl paralleler Aktivitäten aufgetragen.
  - Perioden von Berechnungs- Kommunikations- und Untätigkeitszeiten sind erkennbar.



Beschleunigung und Effizienz

- Beschleunigung (Leistungssteigerung, Speedup):

$$S(n) = \frac{T(1)}{T(n)}$$

- Effizienz:

$$E(n) = \frac{S(n)}{n}$$

- T(1) Ausführungszeit auf einem Einprozessorsystem
- T(n) Ausführungszeit auf einem System mit n Prozessoren

Die „Zeit“ ist auch in Schritte oder Takte messbar.

Skalierbarkeit

Skalierbarkeit eines Parallelrechners

- Das Hinzufügen von weiteren Verarbeitungselementen führt zu einer kürzeren Gesamtausführungszeit, ohne dass das Programm geändert werden muss.
- Wichtig für die Skalierbarkeit sind jeweils angemessene Problemgrößen.
- Bei fester Problemgröße und steigender Prozessorzahl wird ab einer bestimmten Prozessorzahl eine Sättigung eintreten. Die Skalierbarkeit ist in jedem Fall beschränkt (*strong scaling*).
- Darf mit Anzahl an Prozessoren auch die Problemgröße steigen (*weak scaling*), dann muss ein skalierendes Hardware- und Software-System den Sättigungseffekt nicht aufweisen.

Gute Skalierbarkeit:
Lineare Steigerung der Beschleunigung mit einer Effizienz nahe Eins.