(Architetture dei) Calcolatori Elettronici

Prof. Andrea Bartolini
DEI, Università di Bologna
<a.bartolini@unibo.it>

Date importanti per il corso dell'A.A. 2024-25

- Ultimo giorno di lezione:
  - Mercoledi 18 dicembre 2024
- Date degli appelli:
  - Gennaio 2025– scritto + orale
  - Febbraio 2025 – scritto + orale
  - Giugno 2025 – scritto + orale
  - Luglio 2025 – scritto + orale
  - Settembre 2025 – scritto + orale

VISUALIZZAZIONE DI SINTESI

Informatici

Calcolatori Elettronici M

elettronici

Architettura dei Calcolatori Elettronici M

OPZIONALE

Lab of Big Data Architecture

Solo per Elettronici

OPZIONALE

Attività progettuale di Calcolatori Elettronici M

Solo per Informatici

Lezioni in aula + registrazioni

Prova scritta individuale

Voto Max 27/30

Orale / Approfondimento facoltativo

(6 punti)

Progetto

Dimostratore

Report

Presentazione

(da terminare entro 07/2025)

Obiettivo comune è lo studio dei principali aspetti dell'hardware dei calcolatori:

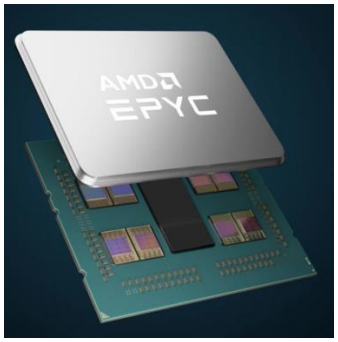
- **Architettura**
- **principi di funzionamento**
- progettazione
- **prestazioni**

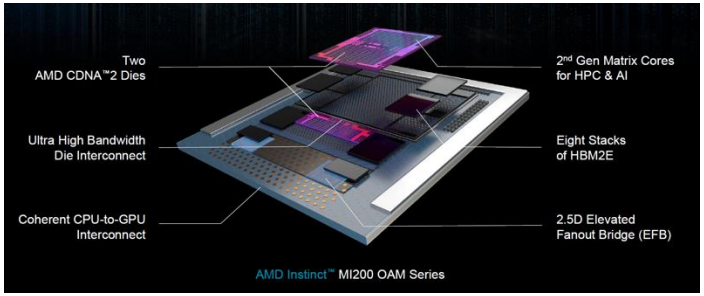
Questa attività si colloca a un livello di astrazione più alto (meno circuitale) rispetto ai contenuti di *calcolatori elettronici T* e degli altri insegnamenti di progettazione digitale già frequentati

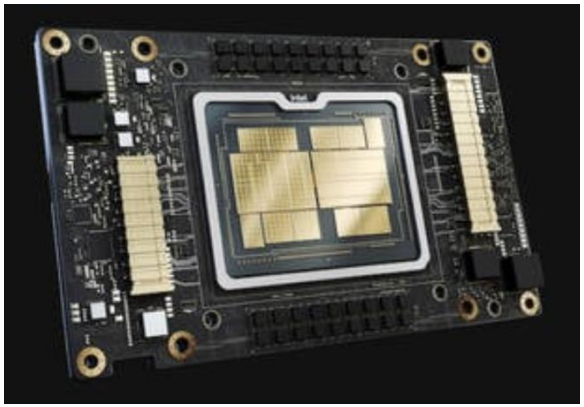
Which car is the best?

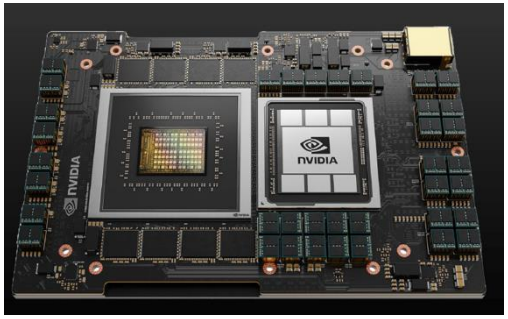
Which car is the best?

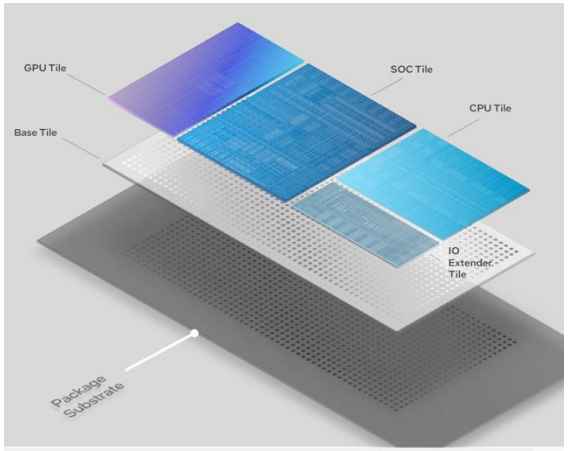
Which processor is the best?



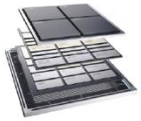


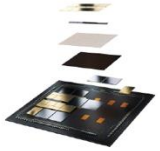


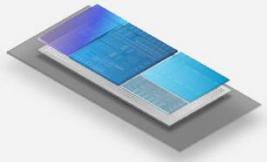




Disaggregation Journey



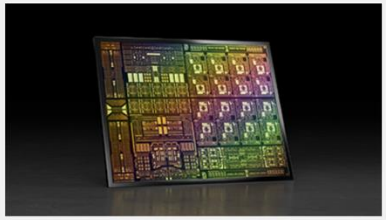




	Kaby Lake G Ultra Thin & Perf Graphics 2017	Lakefield Ultra Thin & Light 2019	Ponte Vecchio High Density & Performance 2022
Architecture	CPU/GFX partitioning	Hybrid Architecture CPU/PCH partitioning	47 Tiles Compute/Memory/IO partitioning
Packaging	2.5D + 2D EMIB + MCP	3D 50µm Foveros	2.5D + 3D EMIB + 36µm Foveros
Process	GloFo 14nm, Intel 14nm	Intel 22FFL, Intel 10nm	TSMC N7, TSMC N5, Intel 7

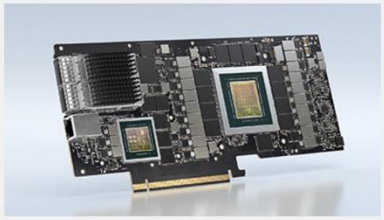
Meteor Lake

Next Step in our Disaggregation Journey



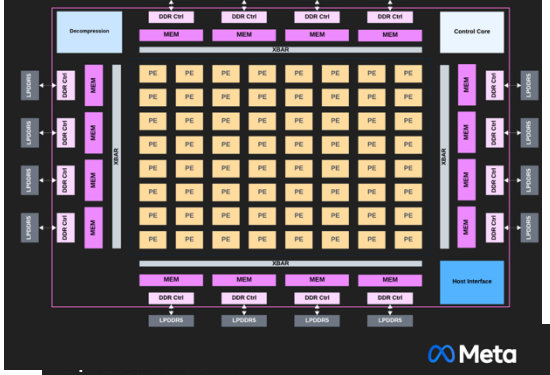
NVIDIA BlueField-3 DPU

NVIDIA BlueField-2 DPU

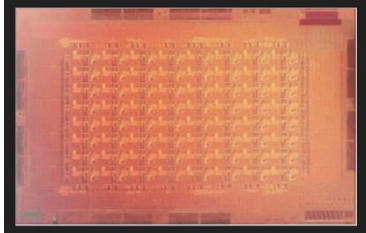


NVIDIA converged accelerators

A new wave…

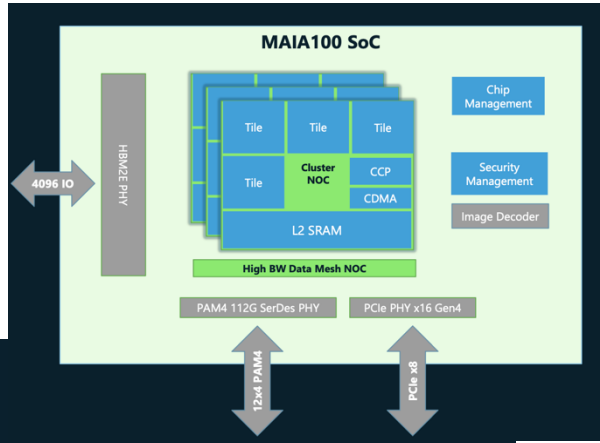


Meta



TECHNOLOGY	TSMC 5nm
FREQUENCY	1.35 GHz
GATE COUNT	2.35B gates, 103M flops
DIMENSIONS	25.6 x 16.4 mm (421 mm2)
PACKAGE	50mm x 40mm
TDP	90 Watts
GEMM TOPS	354 (INT8), 177 (FP16) 2x with sparsity
MEMORY	128GB LPDDR5 6400 BW 204.8GB/s

Meta



Maia 100 Introduction

Microsoft's 1st-gen custom AI Accelerator

- Targets large-scale AI workloads
- Designed specifically for Azure to run production OpenAI models

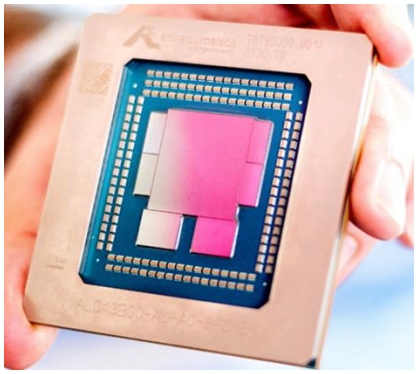
Vertical integration to optimize performance and reduce cost

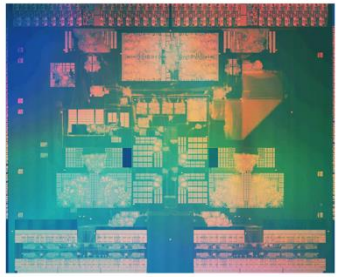
- Software-hardware codesign to unlock new capabilities
- Custom server boards with tailor-made racks
- Improve power efficiency

Maia 100 Specs

Maia 100 Specs

Chip Size	~820mm2@N5
Package/Interposer Technology	TSMC COWOS-S
HBM BW/Cap	1.8TB/s @64GB HBM2E
Peak Dense Tensor POPS	6bit: 3 9bit: 1.5 BF16: 0.8
L1/L2	~500MB
Backend Network BW	600GB/s (12x400gbe)
Host BW (PCIe)	32GB/s PCIe Gen5x8
Design to TDP	700W
Provision TDP	500W





TSMC N7 Process

2.5GHZ CPU

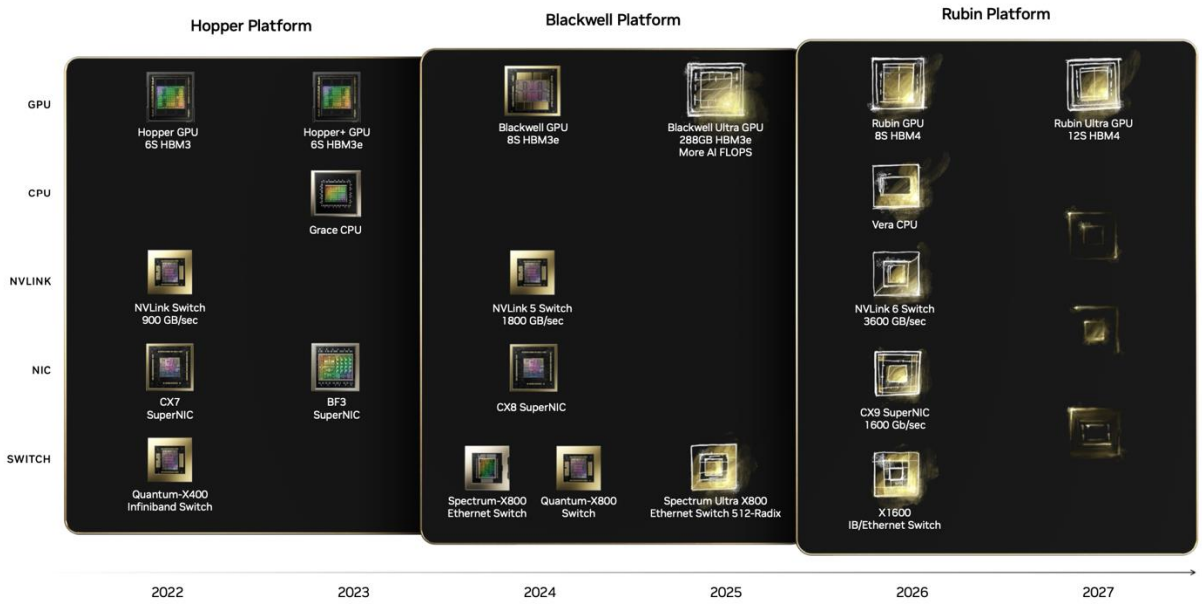
109.9mm²

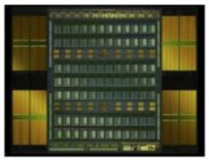
arm

AWS graviton / ARM neoverse n2

Trend #1

Datacenter Scale | One-Year Rhythm | Technology Limits | One Architecture



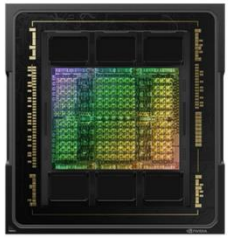


Volta

>21 billion transistors
815mm^2
TSMC 12nm FFN

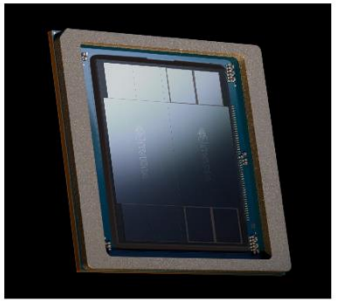
Ampere

>54 billion transistors
826 mm^2
TSMC N7



Hopper

>80 billion transistors
814 mm^2
TSMC 4N

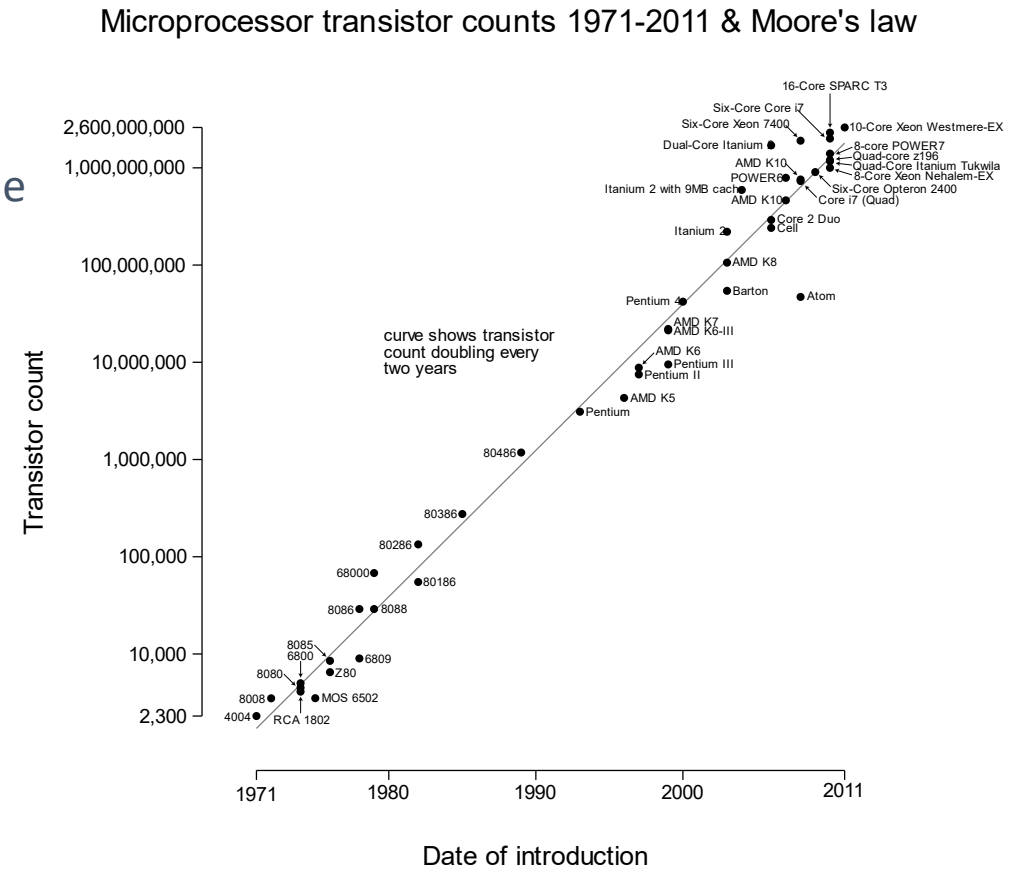


Blackwell

>208 billion transistors
>1600 mm^2
TSMC 4NP

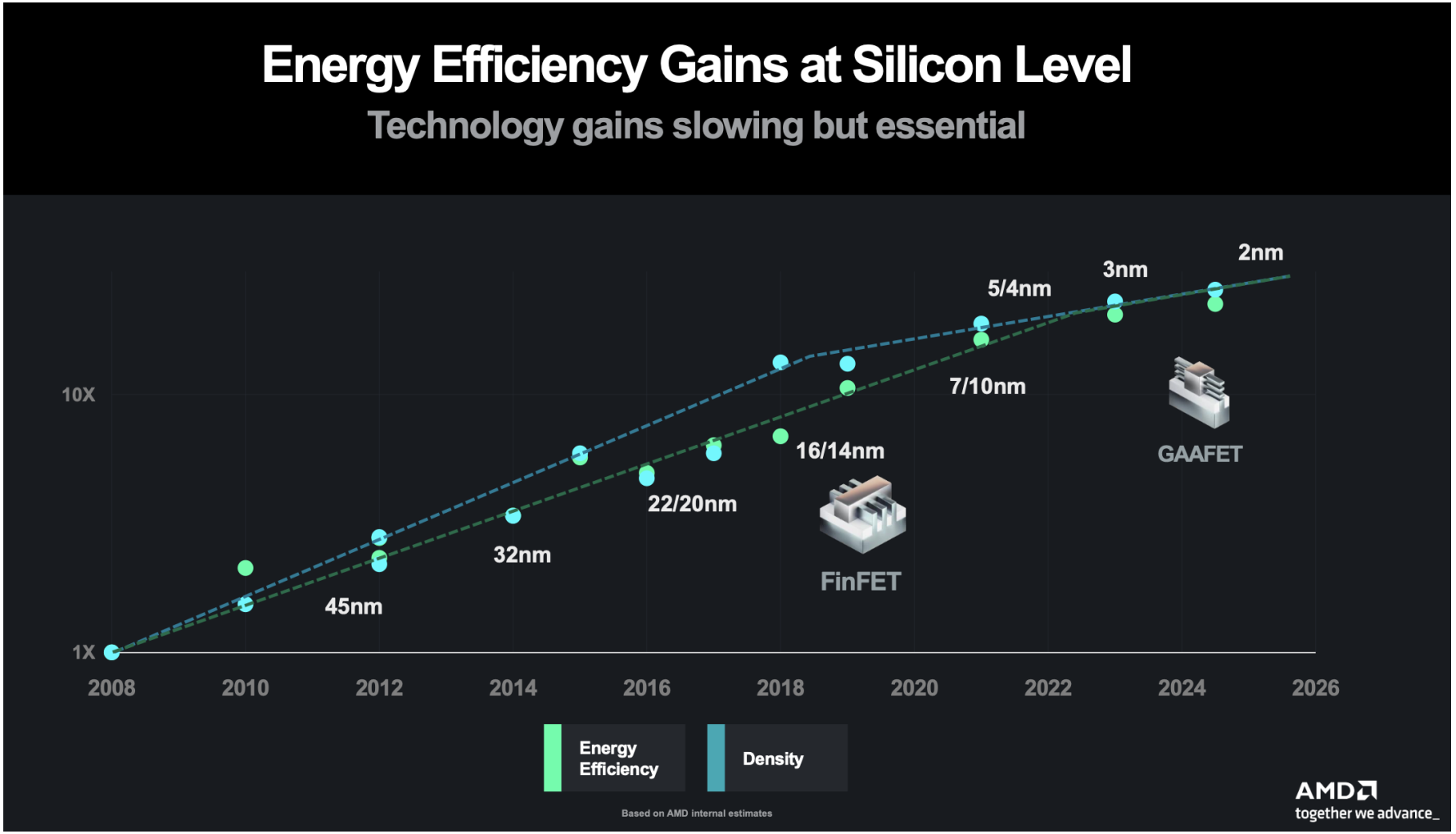
Moore's Law

- Processor transistor budgets grew quickly as microarchitectures became more complex.
- **1985 – Intel 386**
  275K transistors, die size = 43 mm^2
- **2002 – Intel Pentium 4**
  42M transistors, die size = 217 mm^2

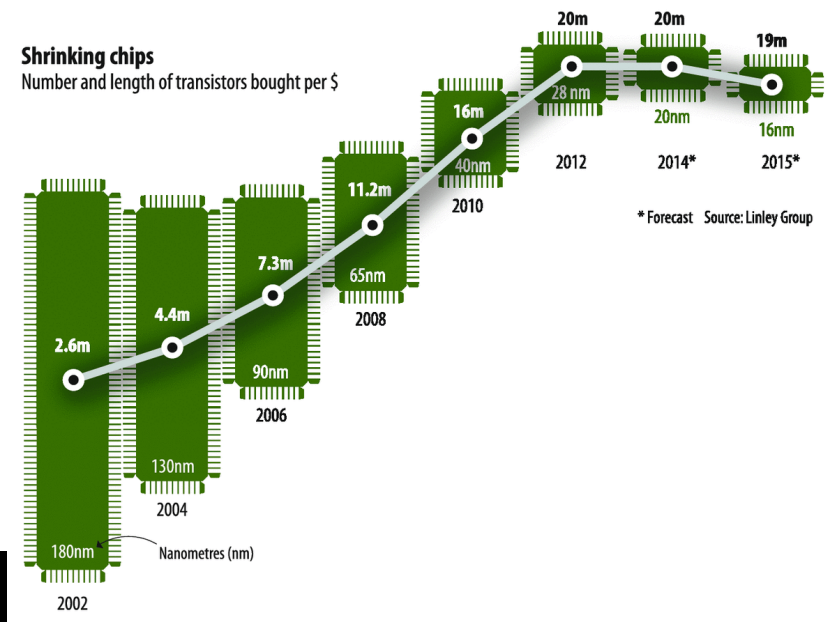


Source: Wgsimon, Wikipedia, CC BY-SA 3.0

Moore’s Law

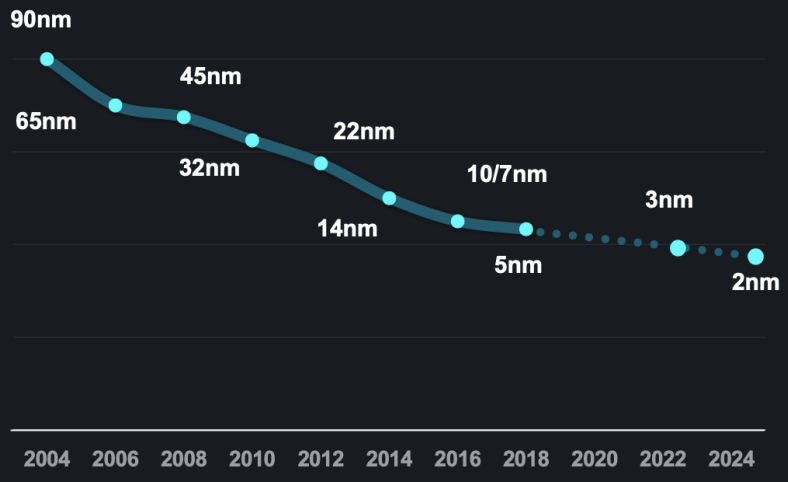

Energy Efficiency Gains at Silicon Level
Technology gains slowing but essential
10X
1X
45nm
32nm
22/20nm
16/14nm
7/10nm
5/4nm
3nm
2nm
FinFET
GAAFET
2008
2010
2012
2014
2016
2018
2020
2022
2024
2026
Energy Efficiency
Density
Based on AMD internal estimates
AMD
together we advance_

Moore's Law



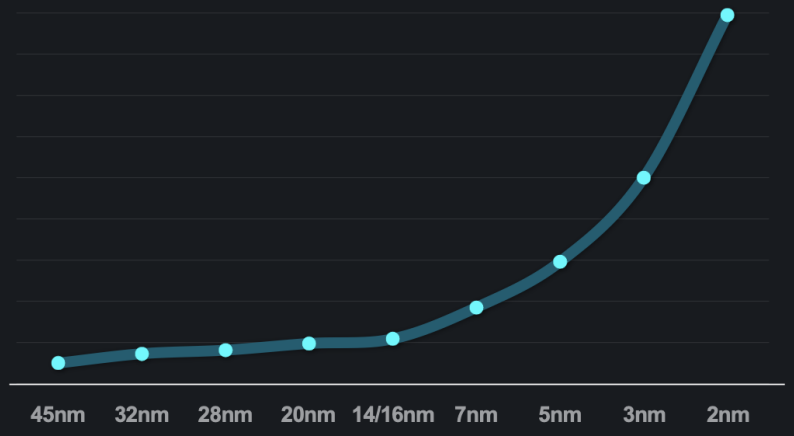
Cost Scaling Challenges

New node introduction rate is slowing



While costs continue to increase

(Cost per yielded mm² for a 250mm² die)



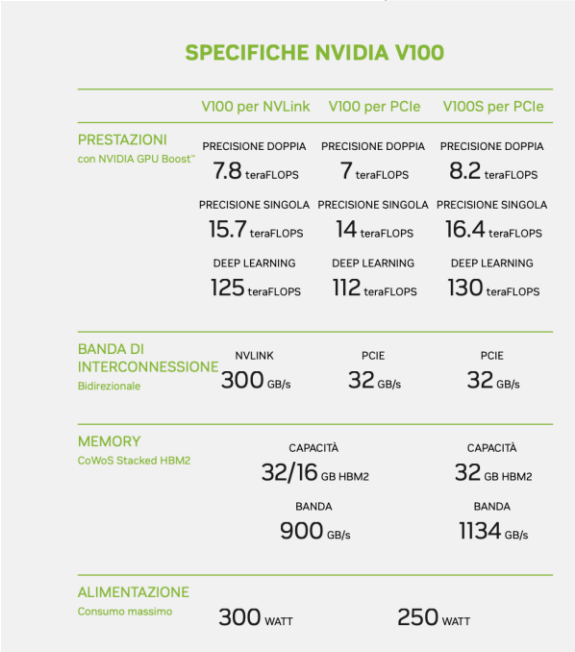
Trend #1 - explained

The "Ampere" A100 silicon has 54 billion transistors crammed into a single 7 nm die (not counting transistor counts of the HBM2E memory stacks).
Y: 2020

	A100 80GB PCIe	A100 80GB SXM
FP64	9,7 TFLOPS	
FP64 Tensor Core	19,5 TFLOPS	
FP32	19,5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
INT8 Tensor Core	624 TOPS | 1248 TOPS*	
Memoria della GPU	HBM2e da 80 GB	HBM2e da 80 GB
Banda di memoria GPU	1.935 GB/s	2.039 GB/s
ermal	300 W	400 W ***

H100 GPU is manufactured on a 'custom version' of TSMC's 4N process, with 80 billion transistors - 68 percent more than the prior-generation 7nm A100 GPU.
Y: 2022

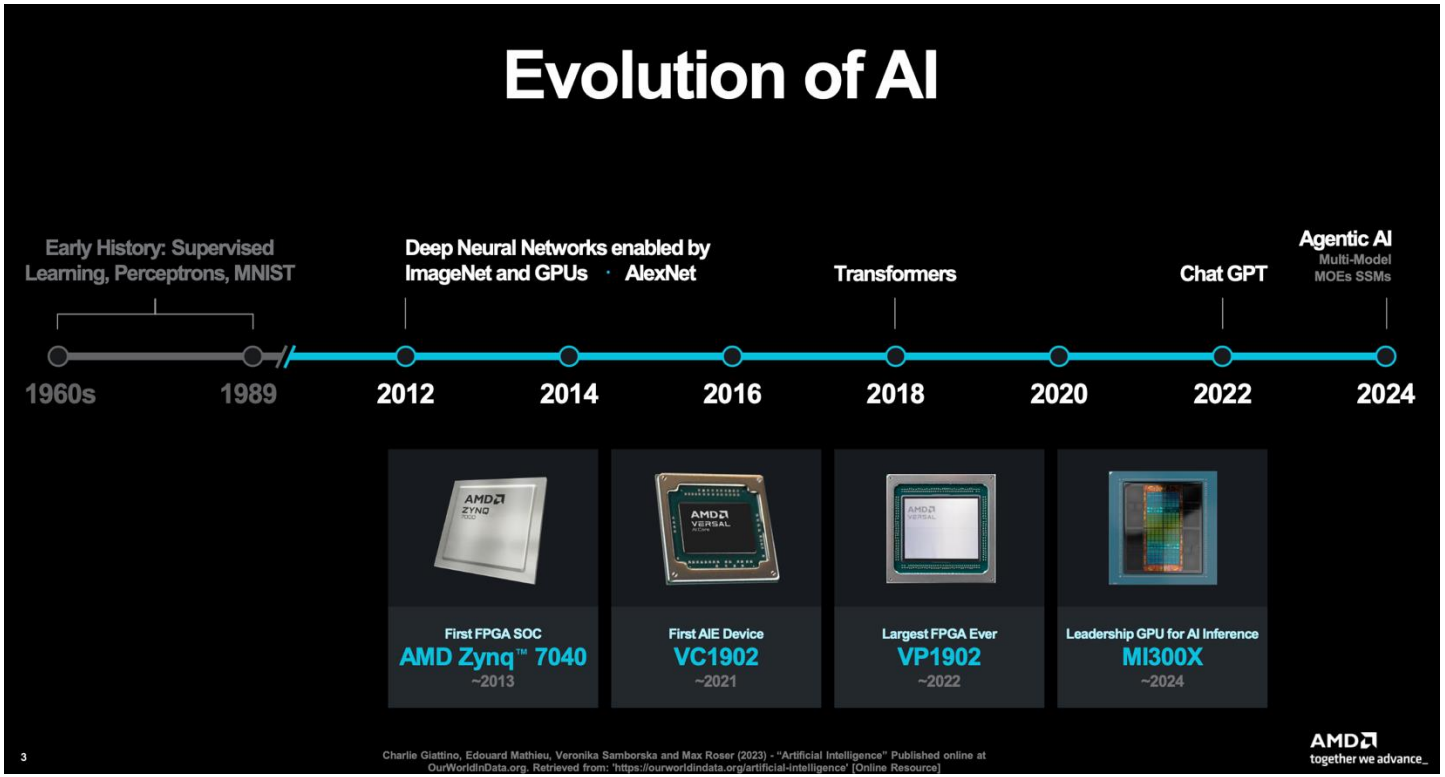
H100 SXM	H100 PCIe
34 teraFLOPS	26 teraFLOPS
67 teraFLOPS	51 teraFLOPS
67 teraFLOPS	51 teraFLOPS
989 teraFLOPS*	756 teraFLOPS*
1979 teraFLOPS*	1.513 teraFLOPS*
1.979 teraFLOPS*	1.513 teraFLOPS*
3.958 teraFLOPS*	3.026 teraFLOPS*
3.958 TOPS*	3.026 TOPS*
80 GB	80 GB
3,35 Tb/s	2 TB/s
7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Fino a 700 W (configurabile)	300-350 W (configurabile)



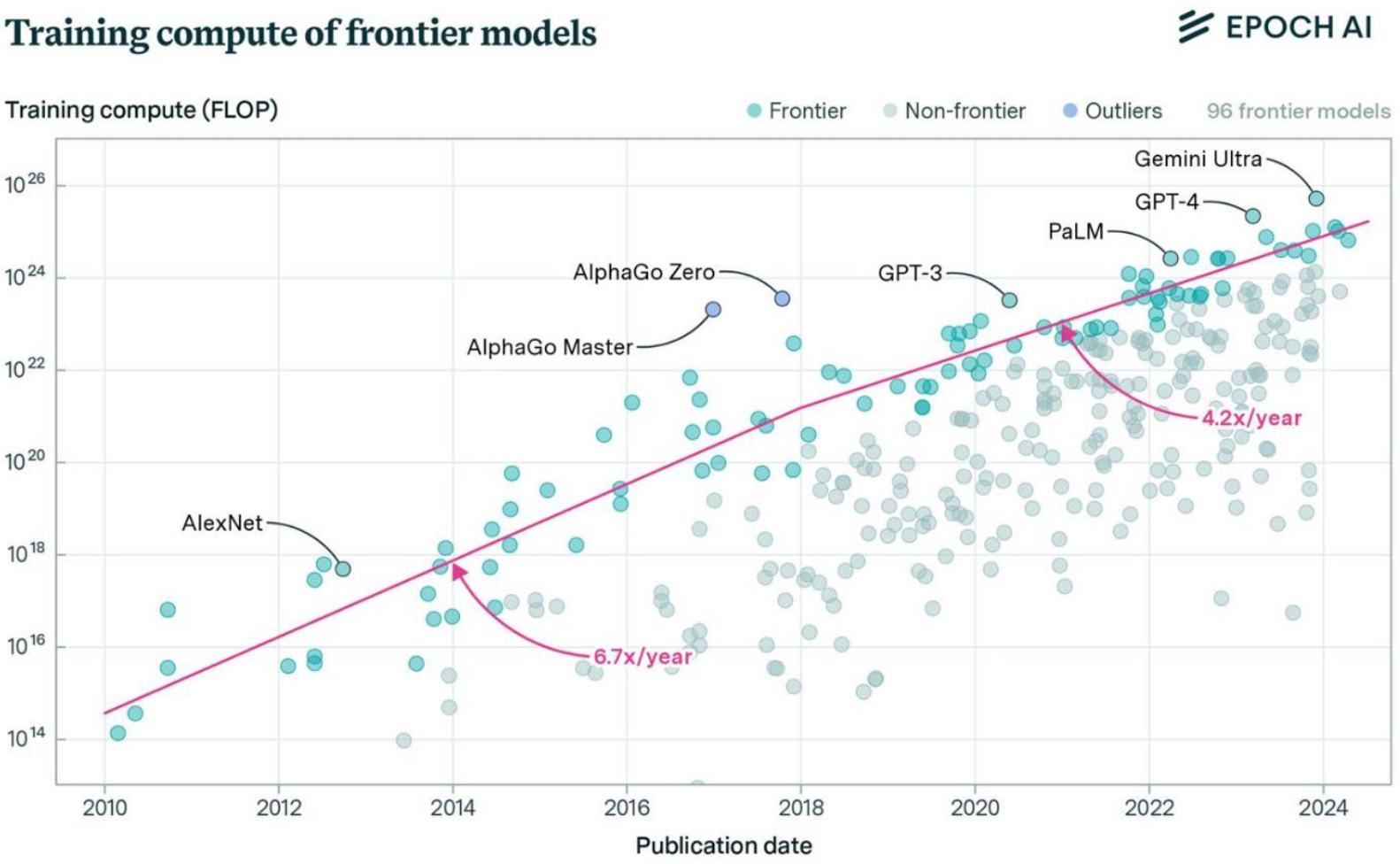
21.1 billion transistors with a die size of 815 mm_2. It is fabricated on a new TSMC 12 nm FFN
Y:2017

Trend #2?

HotChips'24

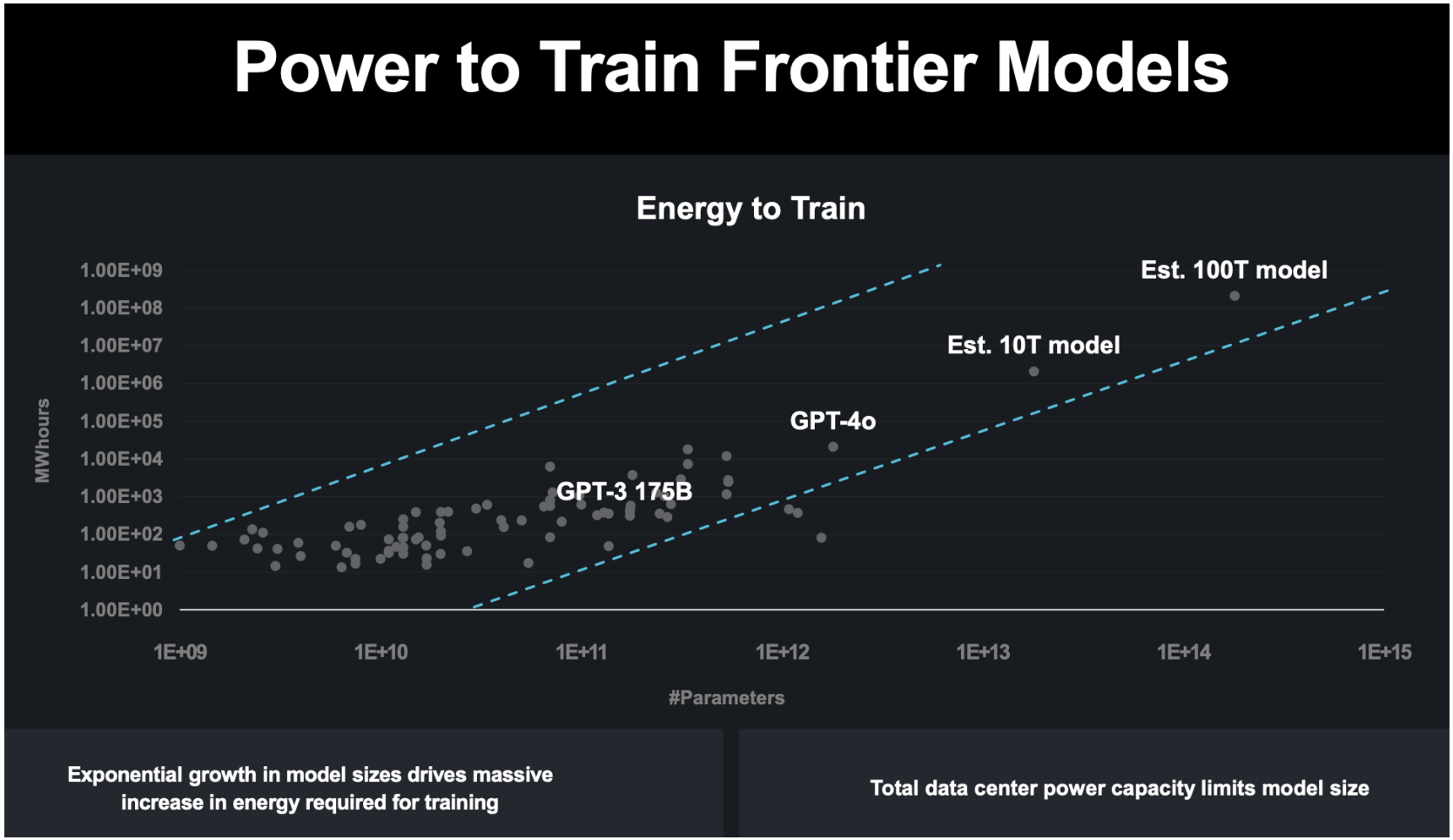


Trend #2?



Training Compute of Frontier AI Models Grows by 4-5x per Year, Sevilla and Roldán (2024)

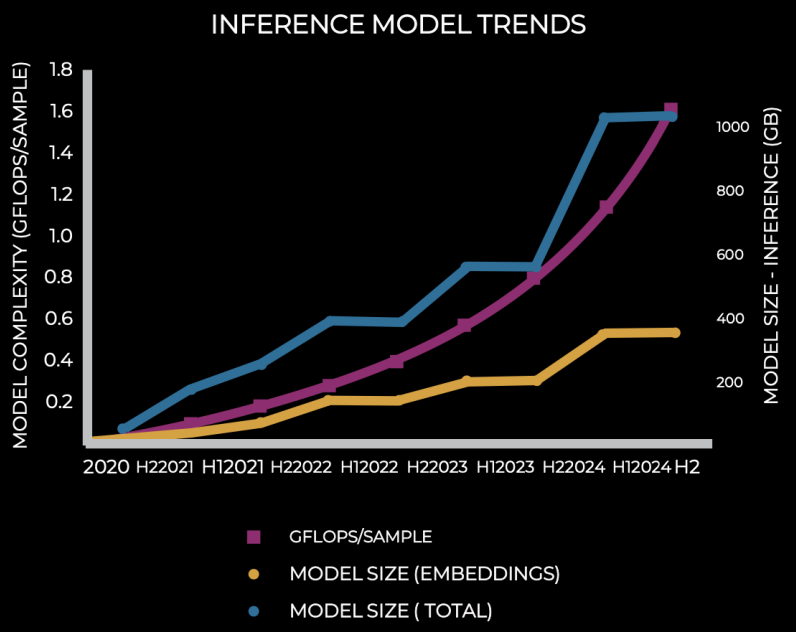
Trend #2?



HotChips'24

Trend #2?

Meta Inference Workload Trends



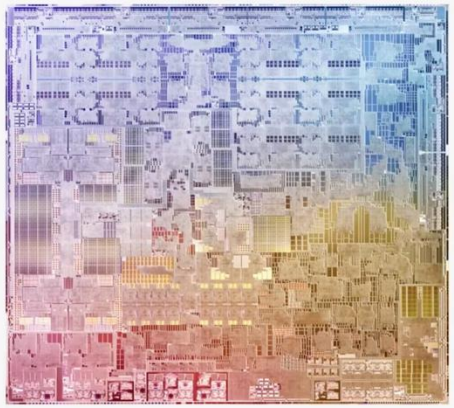
Deep Learning Recommendation Models (DLRM) are increasing in model size (GB) and complexity (GFLOPS)

Models evolved beyond SparseNN for better accuracy and user experience

Emergence of GenAI with LLMs and wide array of models across different use cases

Meta

Trend#2?



M2

Apple M-Series (Vanilla) SoCs			
SoC	**M4**	**M3**	**M2**
CPU Performance	4-core	4-core 16MB Shared L2	4-core (Avalanche) 16MB Shared L2
CPU Efficiency	6-core	4-core 4MB Shared L2	4-core (Blizzard) 4MB Shared L2
GPU	10-Core Same Architecture as M3	10-Core New Architecture - Mesh Shaders & Ray Tracing	10-Core 3.6 TFLOPS
Display Controller	2 Displays?	2 Displays	2 Displays
Neural Engine	16-Core 38 TOPS (INT8)	16-Core 18 TOPS (INT16)	16-Core 15.8 TOPS (INT16)
Memory Controller	LPDDR5X-7500 8x 16-bit CH 120GB/sec Total Bandwidth (Unified)	LPDDR5-6250 8x 16-bit CH 100GB/sec Total Bandwidth (Unified)	LPDDR5-6250 8x 16-bit CH 100GB/sec Total Bandwidth (Unified)
Max Memory Capacity	24GB?	24GB	24GB
Encode/ Decode	8K H.264, H.265, ProRes, ProRes RAW, AV1 (Decode)	8K H.264, H.265, ProRes, ProRes RAW, AV1 (Decode)	8K H.264, H.265, ProRes, ProRes RAW
USB	USB4/Thunderbolt 3 ? Ports	USB4/Thunderbolt 3 2x Ports	USB4/Thunderbolt 3 2x Ports
Transistors	28 Billion	25 Billion	20 Billion
Mfc. Process	TSMC N3E	TSMC N3B	TSMC N5P

Matrix Engines

Matrix Engines are primarily designed to accelerate GEMM / GEMV

- All have slightly different implementations and ways of using them

Many modern GPUs now have Tensor / Matrix Cores:

- NVIDIA **Tensor Cores** since V100 in 2017
- Intel **Xe Cores** on Ponte Vecchio
- AMD **Matrix Cores** on MI100, MI210 and MI250X

Many HPC and consumer CPUs have matrix engines too:

- Intel Advanced Matrix eXtension (AMX) in Sapphire Rapids
- IBM Matrix Multiply Assist (MMA) in Power10
- Apple Advanced Matrix eXtension (AMX) in M1-M3
- **Arm Scalable Matrix Extension (SME)** in Apple M4

SME features

The following SME features are reported for Apple M4

- `FEAT_SME`
- `FEAT_SME2`
- `SME_F32F32`
- `SME_BI32I32`
- `SME_B16F32`
- `SME_F16F32`
- `SME_I8I32`
- `SME_I16I32`
- `FEAT_SME_F64F64`
- `FEAT_SME_I16I64`

Teaser #2

Enabling Further Efficiency through Advanced Quantization

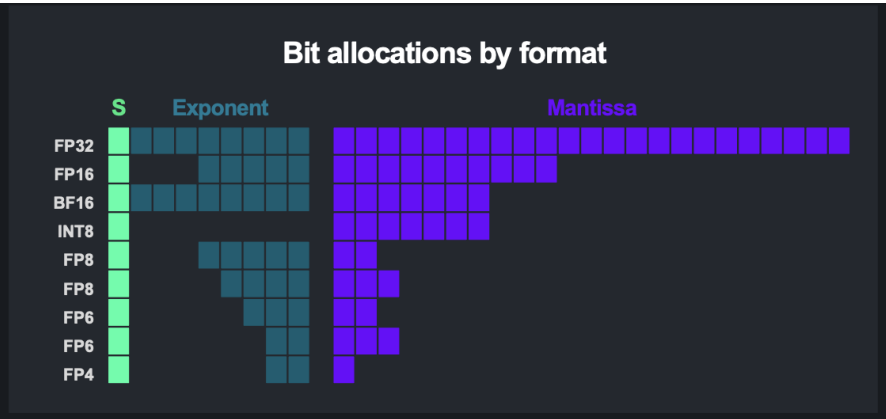
Adapt algorithms to use lower precision math for significant improvements in energy efficiency

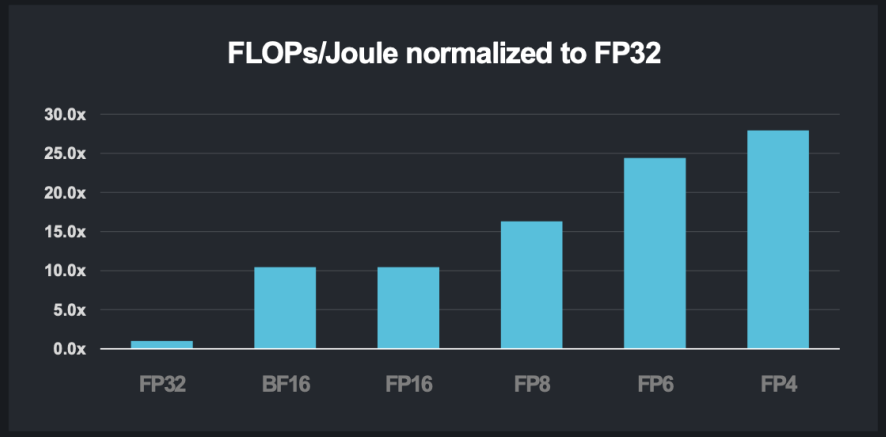
- FP8 in MI325X and FP4 and FP6 in MI350X

Advancing innovation in quantization

Novel research into accumulator-aware quantization[1]

Collaborating through open-source[2]





[1] https://arxiv.org/abs/2301.13376
[2] *brevitas/src/brevitas_examples/imagenet_classification/a2q

Source: Internal AMD Data

AMD
together we advance_

LLAMA 3.1 Training Cost

- Meta Llama 3 405B pre-trained using 3.8×10^{25} FLOPs, almost 50× more than the largest version of Llama 2
- Pre-trained using scaling laws tuned with ”small” pre-training models ranging from 40M to 15B parameters w. compute budgets between 6×10^{18} FLOPs and 10^{22} FLOPs.
- Llama 3 405B is trained on up to 16K H100 GPUs, each running at 700W TDP with 80GB HBM3, using Meta’s Grand Teton AI server platform.

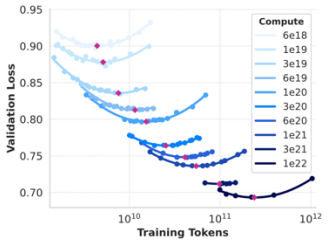


Figure 2 Scaling law IsoFLOPs curves between 6×10^{18} and 10^{22} FLOPs. The loss is the negative log-likelihood on a held-out validation set. We approximate measurements at each compute scale using a second degree polynomial.

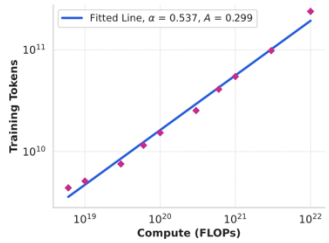


Figure 3 Number of training tokens in identified compute-optimal models as a function of pre-training compute budget. We include the fitted scaling-law prediction as well. The compute-optimal models correspond to the parabola minimums in Figure 2.

Technical Specifications	H100 SXM
FP64	34 teraFLOPS
FP64 Tensor Core	67 teraFLOPS
FP32	67 teraFLOPS
TF32 Tensor Core*	989 teraFLOPS
BFLOAT16 Tensor Core*	1,979 teraFLOPS
FP16 Tensor Core*	1,979 teraFLOPS
FP8 Tensor Core*	3,958 teraFLOPS
INT8 Tensor Core*	3,958 TOPS
GPU Memory	80GB
GPU Memory Bandwidth	3.35TB/s
Decoders	7 NVDEC 7 JPEG
Max Thermal Design Power (TDP)	Up to 700W (configurable)
Multi-Instance GPUs	Up to 7 MIGs @ 10GB each
Form Factor	SXM

16K * ~1PFLOP/s @TF32
~16 ExaFLOP/s @TF32
~ 6.5 ExaFLOP/s @TF32

~1.21x Eagle

GPUs	TP	CP	PP	DP	Seq. Len.	Batch size/DP	Tokens/Batch	TFLOPs/GPU	BF16 MFU
8,192	8	1	16	64	8,192	32	16M	430	43%
16,384	8	1	16	128	8,192	16	16M	400	41%
16,384	8	16	16	4	131,072	16	16M	380	38%

Table 4 Scaling configurations and MFU for each stage of Llama 3 405B pre-training. See text and Figure 5 for descriptions of each type of parallelism.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure	2,073,600	561.20	846.84	

Power….

One interesting observation is the impact of environmental factors on training performance at scale. For Llama 3 405B , we noted a diurnal 1-2% throughput variation based on time-of-day. This fluctuation is the result of higher mid-day temperatures impacting GPU dynamic voltage and frequency scaling.

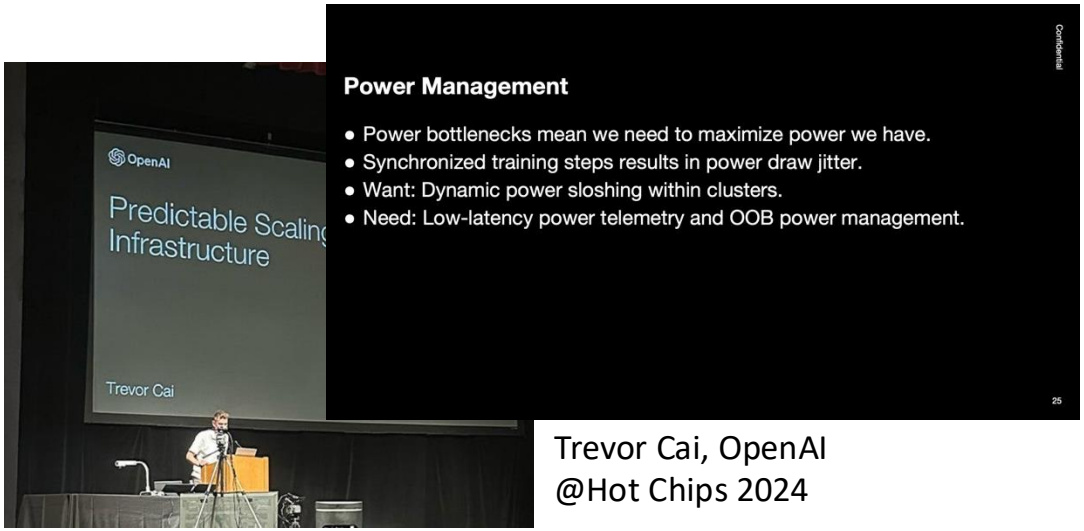
During training, tens of thousands of GPUs may increase or decrease power consumption at the same time, for example, due to all GPUs waiting for checkpointing or collective communications to finish, or the startup or shutdown of the entire training job. When this happens, it can result in instant fluctuations of power consumption across the data center on the order of tens of megawatts, stretching the limits of the power grid. This is an ongoing challenge for us as we scale training for future, even larger Llama models.

Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta[1]
[1]A detailed contributor list can be found in the appendix of this paper.

https://arxiv.org/abs/2407.21783



Trevor Cai, OpenAI
@Hot Chips 2024

Component	Category	Interruption Count	% of Interruptions
Faulty GPU	GPU	148	30.1%
GPU HBM3 Memory	GPU	72	17.2%
Software Bug	Dependency	54	12.9%
Network Switch/Cable	Network	35	8.4%
Host Maintenance	Unplanned Maintenance	32	7.6%
GPU SRAM Memory	GPU	19	4.5%
GPU System Processor	GPU	17	4.1%
NIC	Host	7	1.7%
NCCL Watchdog Timeouts	Unknown	7	1.7%
Silent Data Corruption	GPU	6	1.4%
GPU Thermal Interface + Sensor	GPU	6	1.4%
SSD	Host	3	0.7%
Power Supply	Host	3	0.7%
Server Chassis	Host	2	0.5%
IO Expansion Board	Host	2	0.5%
Dependency	Dependency	2	0.5%
CPU	Host	2	0.5%
System Memory	Host	2	0.5%

Table 5 Root-cause categorization of unexpected interruptions during a 54-day period of Llama 3 405B pre-training. About 78% of unexpected interruptions were attributed to confirmed or suspected hardware issues.

The End of Dennard's scaling

new VLSI gen.

old VLSI gen.

$L' = L / 2$

$V' = V / 2$

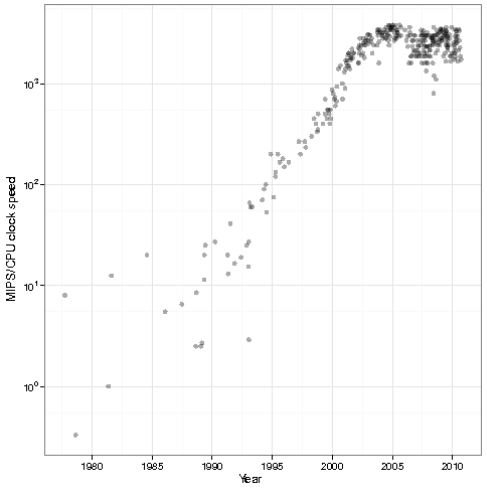
$F' = F * 2$

do not hold anymore!

$D' = 1 / L^2 = 4D$

$P' = P$

The core frequency and performance do not grow following the Moore's law any longer



- Now, power and/or heat generation are the limiting factors of the down-scaling
- Supply voltage reduction is becoming difficult, because Vth cannot be decreased any more, as described later.
- Growth rate in clock frequency and chip area becomes smaller.

$L' = L / 2$

$V' = \sim V$

$F' = \sim F * 2$

$D' = 1 / L^2 = 4 * D$

$P' = 4 * P$

Increase the number of cores to maintain the architectures evolution on the Moore's law

The power crisis!

Programming crisis!

C.Cavazzoni

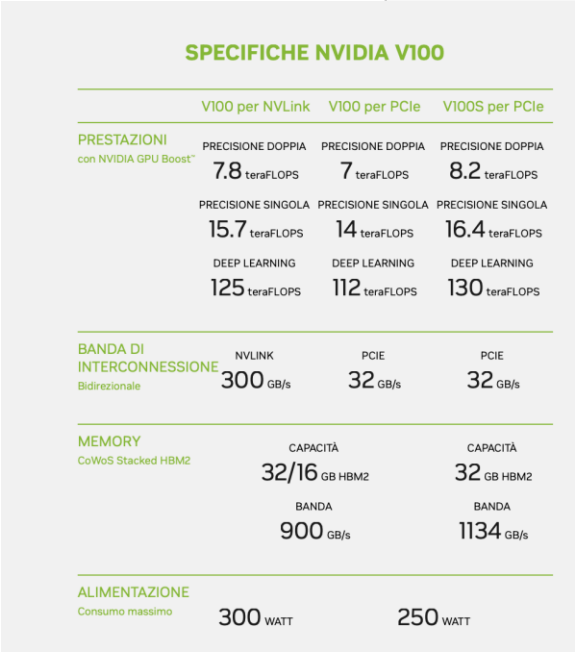
… in practice

The "Ampere" A100 silicon has 54 billion transistors crammed into a single 7 nm die (not counting transistor counts of the HBM2E memory stacks).
Y: 2020

	A100 80GB PCIe	A100 80GB SXM
FP64	9,7 TFLOPS	
FP64 Tensor Core	19,5 TFLOPS	
FP32	19,5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
INT8 Tensor Core	624 TOPS | 1248 TOPS*	
Memoria della GPU	HBM2e da 80 GB	HBM2e da 80 GB
Banda di memoria GPU	1.935 GB/s	2.039 GB/s
ermal	300 W	400 W ***

H100 GPU is manufactured on a 'custom version' of TSMC's 4N process, with 80 billion transistors - 68 percent more than the prior-generation 7nm A100 GPU.
Y: 2022

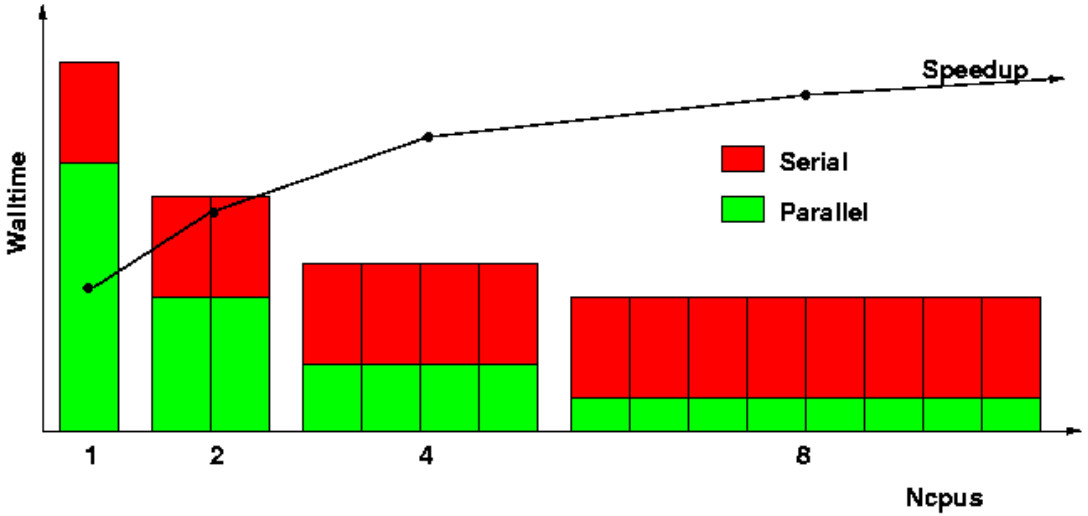
H100 SXM	H100 PCIe
34 teraFLOPS	26 teraFLOPS
67 teraFLOPS	51 teraFLOPS
67 teraFLOPS	51 teraFLOPS
989 teraFLOPS*	756 teraFLOPS*
1979 teraFLOPS*	1.513 teraFLOPS*
1.979 teraFLOPS*	1.513 teraFLOPS*
3.958 teraFLOPS*	3.026 teraFLOPS*
3.958 TOPS*	3.026 TOPS*
80 GB	80 GB
3,35 Tb/s	2 TB/s
7 NVDEC 7 JPEG	7 NVDEC 7 JPEG
Fino a 700 W (configurabile)	300-350 W (configurabile)



21.1 billion transistors with a die size of 815 mm_2. It is fabricated on a new TSMC 12 nm FFN
Y:2017

Adam's Law (a.k.a. Diminishing returns)

In a massively parallel context, an upper limit for the scalability of parallel applications is determined by the fraction of the overall execution time spent in non-scalable operations (Amdahl's law).



maximum speedup tends to
$1 / (1 - P)$
P= parallel fraction

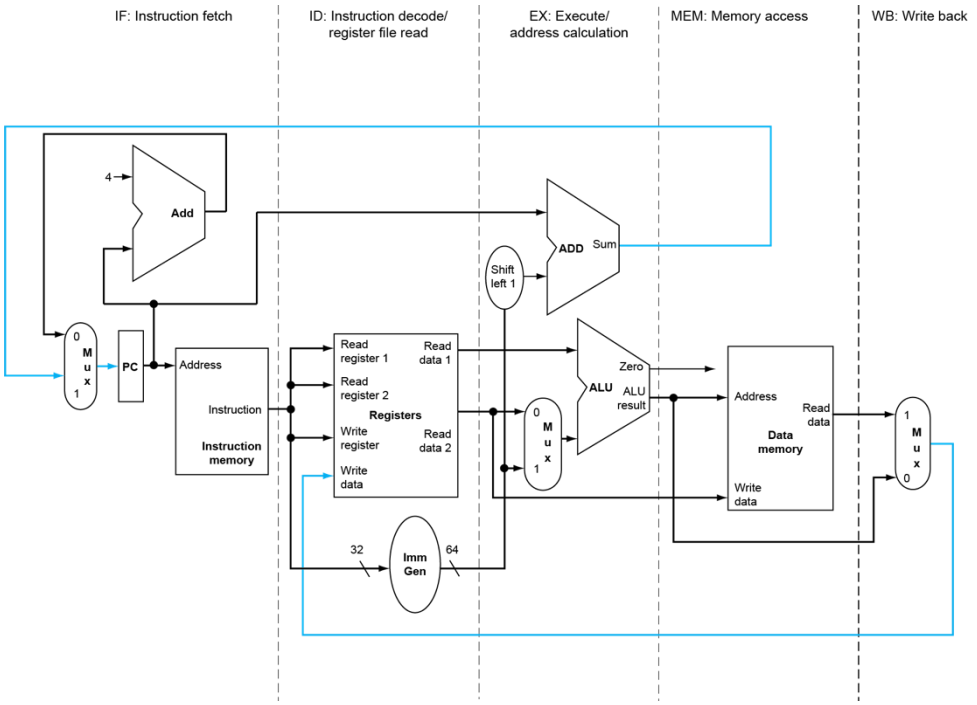
1 000 000 core

P = 0.999 999

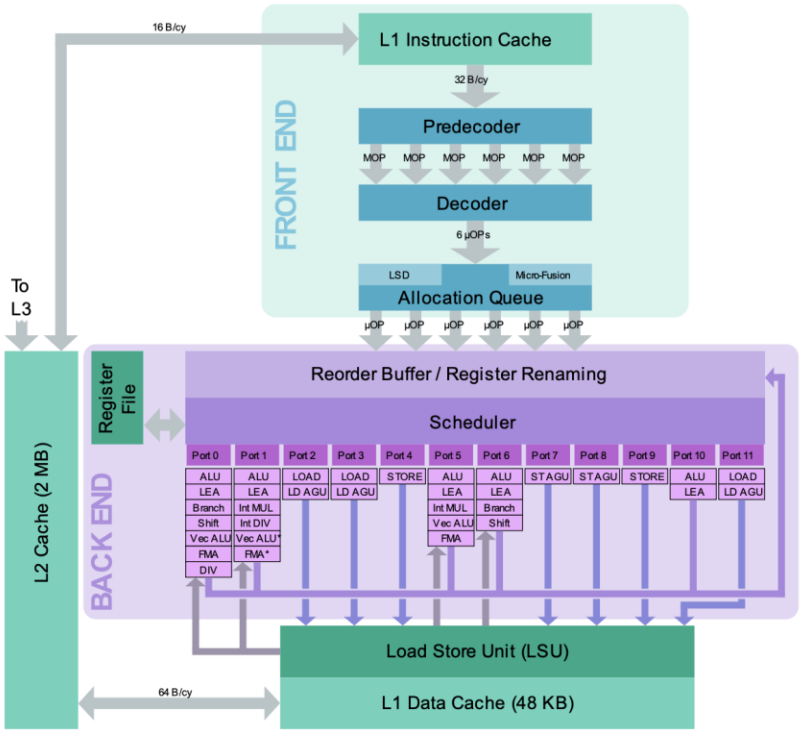
serial fraction= 0.000001

Proportionally smaller profits or benefits derived from something as more money or energy is invested in it. (Diminishing returns).

From in-order to superscalar non-blocking



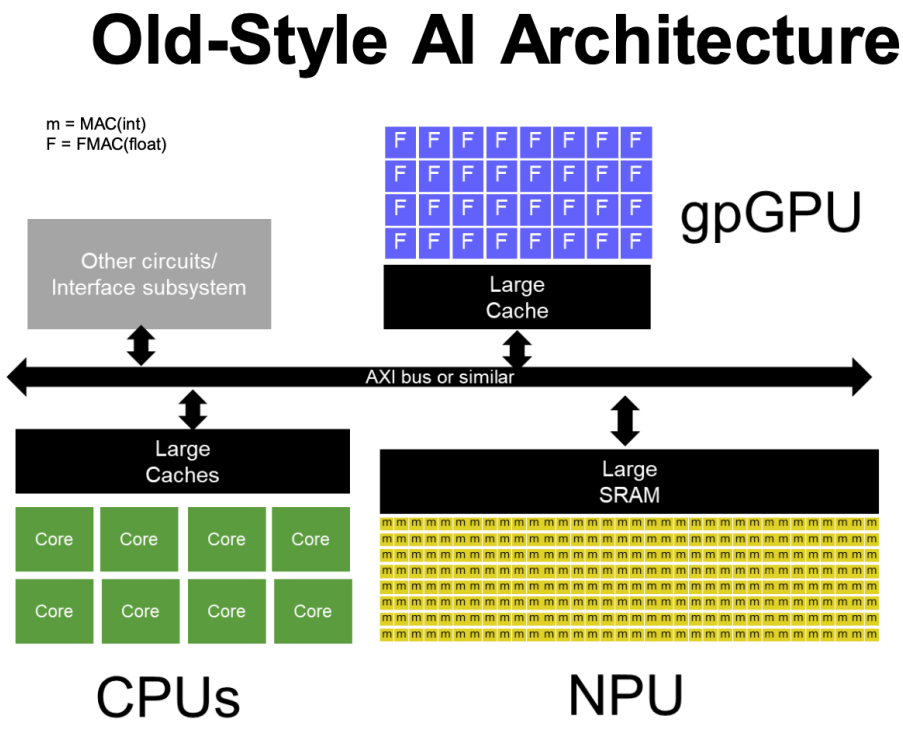
Intel Saffire Rapids Pipeline



Specialization

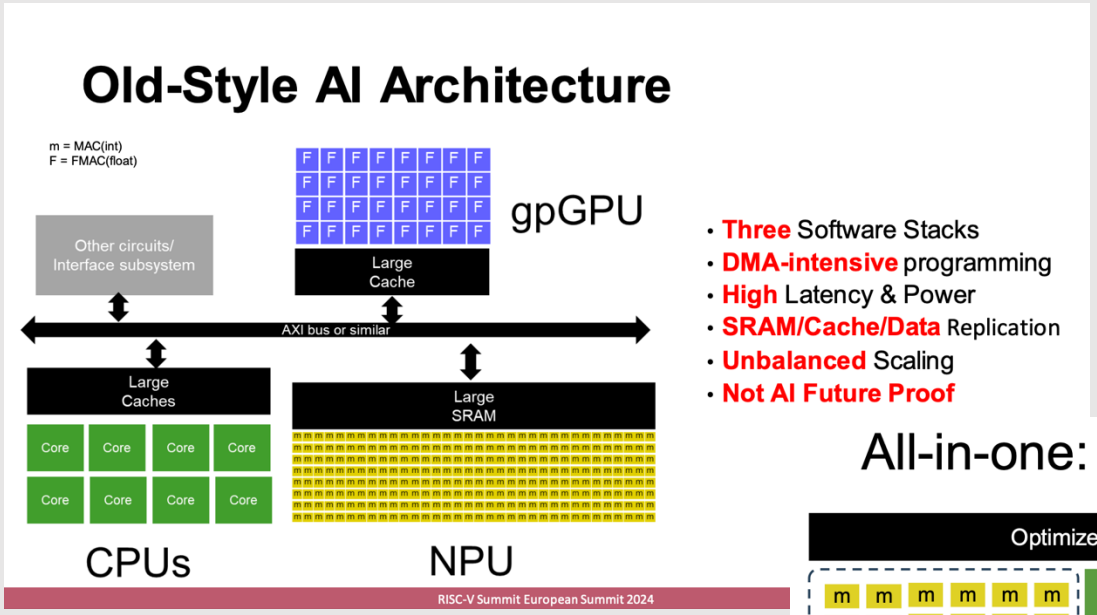
- Today, we often need to look beyond general-purpose programmable processors to meet our design goals.
- We trade flexibility for efficiency.
- We remove the ability to run all programs and design for a narrow workload, perhaps even a single algorithm.
- These “accelerators” can be 10-1000x better than a general-purpose solution in terms of power and performance.

What about GenAI?

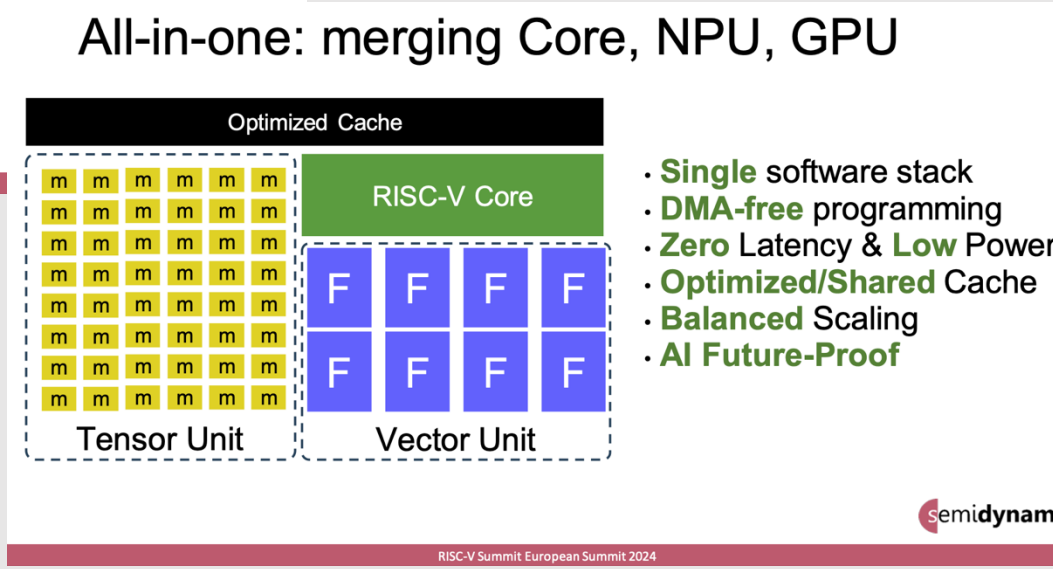


- **Three** Software Stacks
- **DMA-intensive** programming
- **High** Latency & Power
- **SRAM/Cache/Data** Replication
- **Unbalanced** Scaling
- **Not AI Future Proof**

What about GenAI?



- **Three** Software Stacks
- **DMA-intensive** programming
- **High** Latency & Power
- **SRAM/Cache/Data** Replication
- **Unbalanced** Scaling
- **Not AI Future Proof**

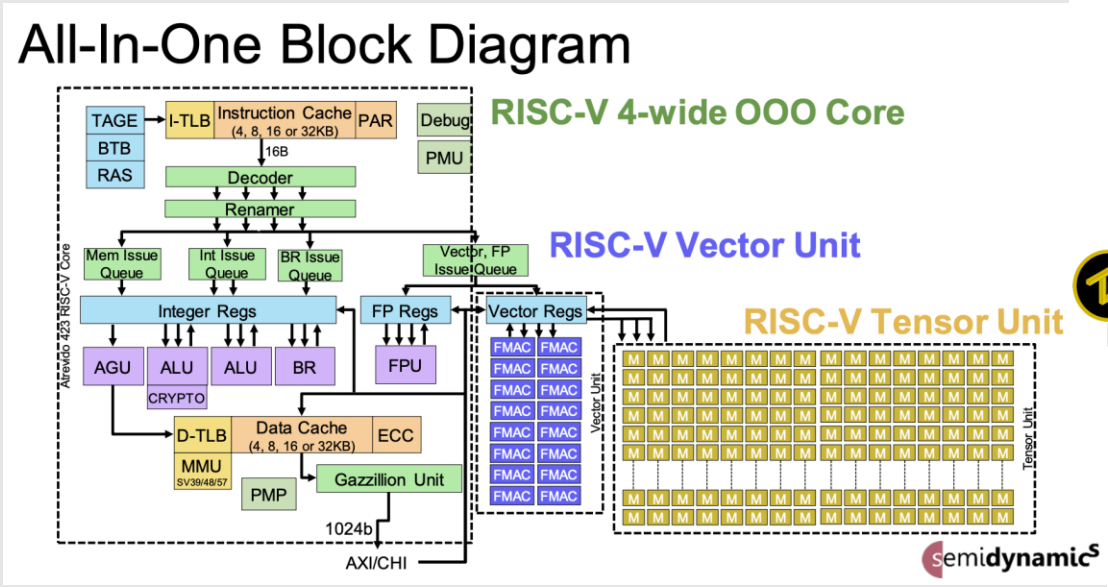


- **Single** software stack
- **DMA-free** programming
- **Zero** Latency & **Low** Power
- **Optimized/Shared** Cache
- **Balanced** Scaling
- **AI Future-Proof**

semidynamics

Courtesy of Roger Espasa,
RISC-V SUMMIT EU 24

What about GenAI?



We'll use our 1 $TOPS_8$ T1 Tensor Unit…

Product	T1	T2	T4	T8
MACs	512	1024	2048	4096
Local SRAM?	No	No	64KB	128KB
INT8 TOPS/GHz	1	2	4	8
INT16 TOPS/GHz	0.5	1	2	4
BF16 TOPS/GHz	0.5	1	2	4
FP16 TOPS/GHz	0.5	1	2	4

semidynamics

We'll use our 128 $GOPS_8$ V128 Vector Unit…

Product	V128	V256	V512
FMACs	8	16	32
INT8 GOPS/GHz	128	256	512
INT16 GOPS/GHz	64	128	256
BF16 GOPS/GHz	64	128	256
FP16 GOPS/GHz	64	128	256
FP32 GOPS/GHz	32	64	128
FP64 GOPS/GHz	16	32	64

semidynamics

Courtesy of Roger Espasa,
RISC-V SUMMIT EU 24

What about GenAI?

Llama-2
FP16,
7B params

Operators	Scalar	T1	T1+V128
Matmul	99%	20%	55%
Activations	1%	80%	45%
Concat	0.11%	19%	17%
Sigmoid	0.09%	16%	2%
ScatterND	0.09%	15%	17%
Div	0.06%	9.5%	2%
Mul	0.03%	5.7%	2.4%
Slice	0.03%	5.0%	1.3%
Exp	0.03%	4.4%	0.5%
Other	0.54%	5.4%	2.8%
Speedup	1X	170X	470X

semidynamics

Courtesy of Roger Espasa,
RISC-V SUMMIT EU 24

IL MODELLO DI RIFERIMENTO del calcolatore

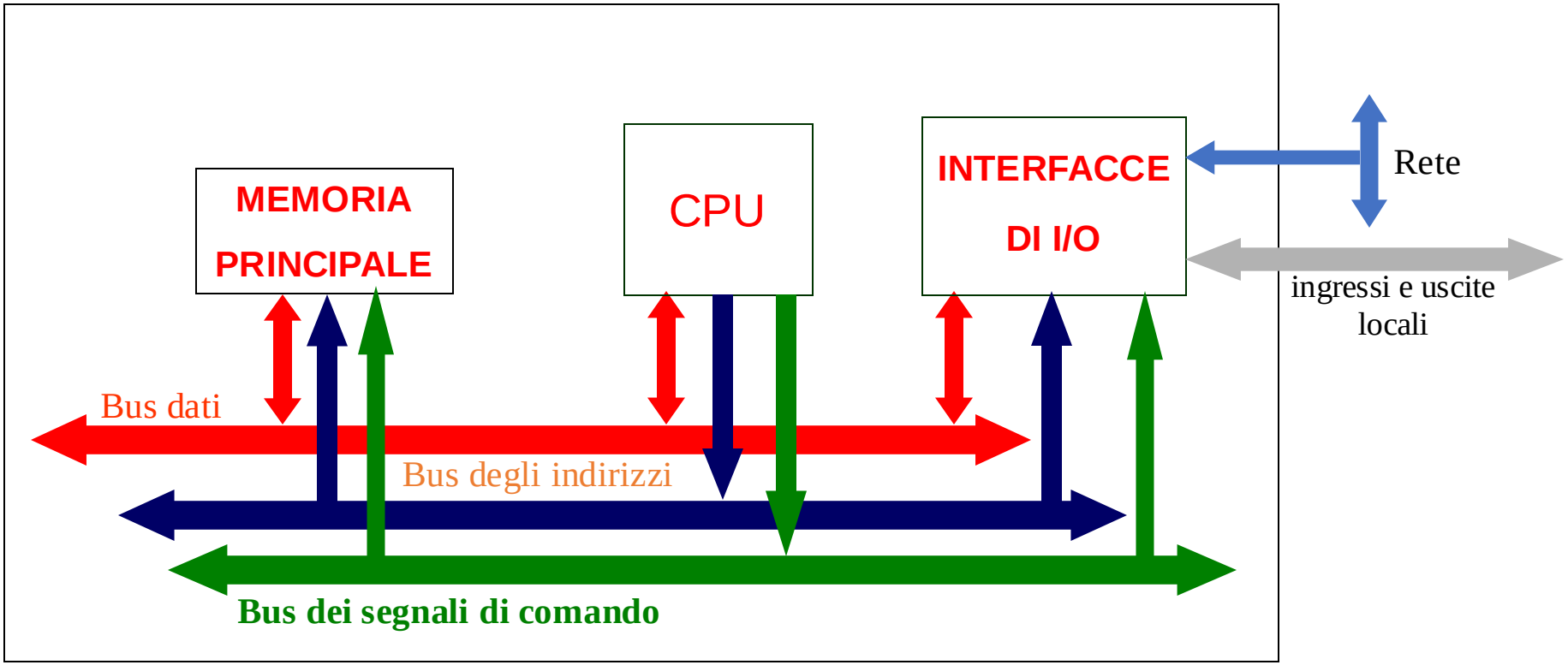
Resta quello visto in Calcolatori Elettronici T:

Macchina digitale
a
esecuzione sequenziale
e
programma memorizzato
(Von Neumann, 1940)

Ma ne verranno considerate delle realizzazioni più generali e molto più potenti, rese possibili dal progresso della tecnologia e adeguate alle applicazioni citate nei lucidi precedenti

Struttura di un calcolatore

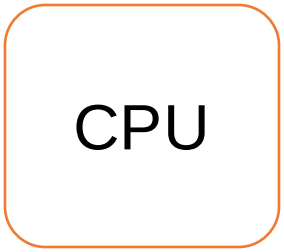
Bus dati, bus degli indirizzi e bus dei segnali di comando



In questo schema a blocchi la CPU genera i segnali di indirizzo e di comando per la memoria e le interfacce

Struttura di un calcolatore moderno

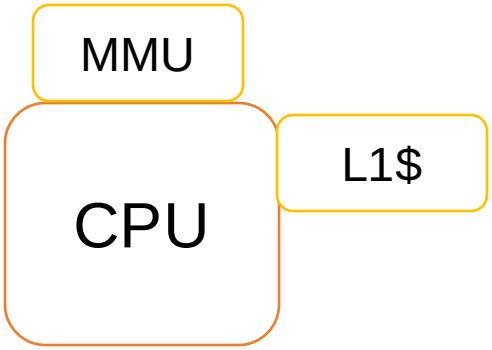
System on Chip Desing, many-core, SIMD/Vettoriale, integrazione 2.5D



1. Ripasso Pipeline
- ISA RISC-V

Struttura di un calcolatore moderno

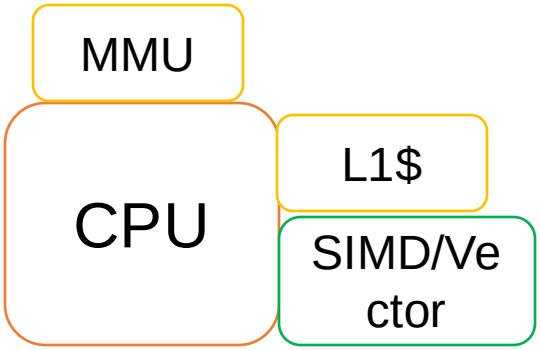
System on Chip Desing, many-core, SIMD/Vettoriale, integrazione 2.5D



2. Gerarchia di memorie
- Cache
- Memoria Virtuale

Struttura di un calcolatore moderno

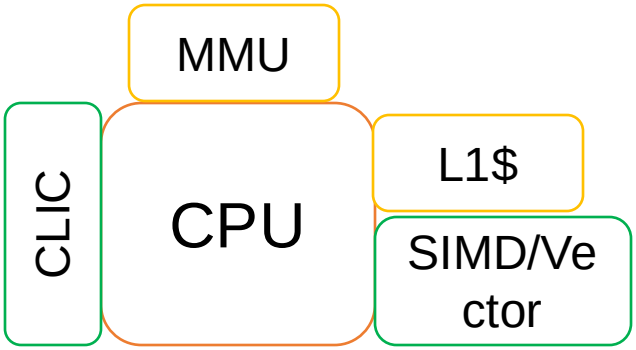
System on Chip Desing, many-core, SIMD/Vettoriale, integrazione 2.5D

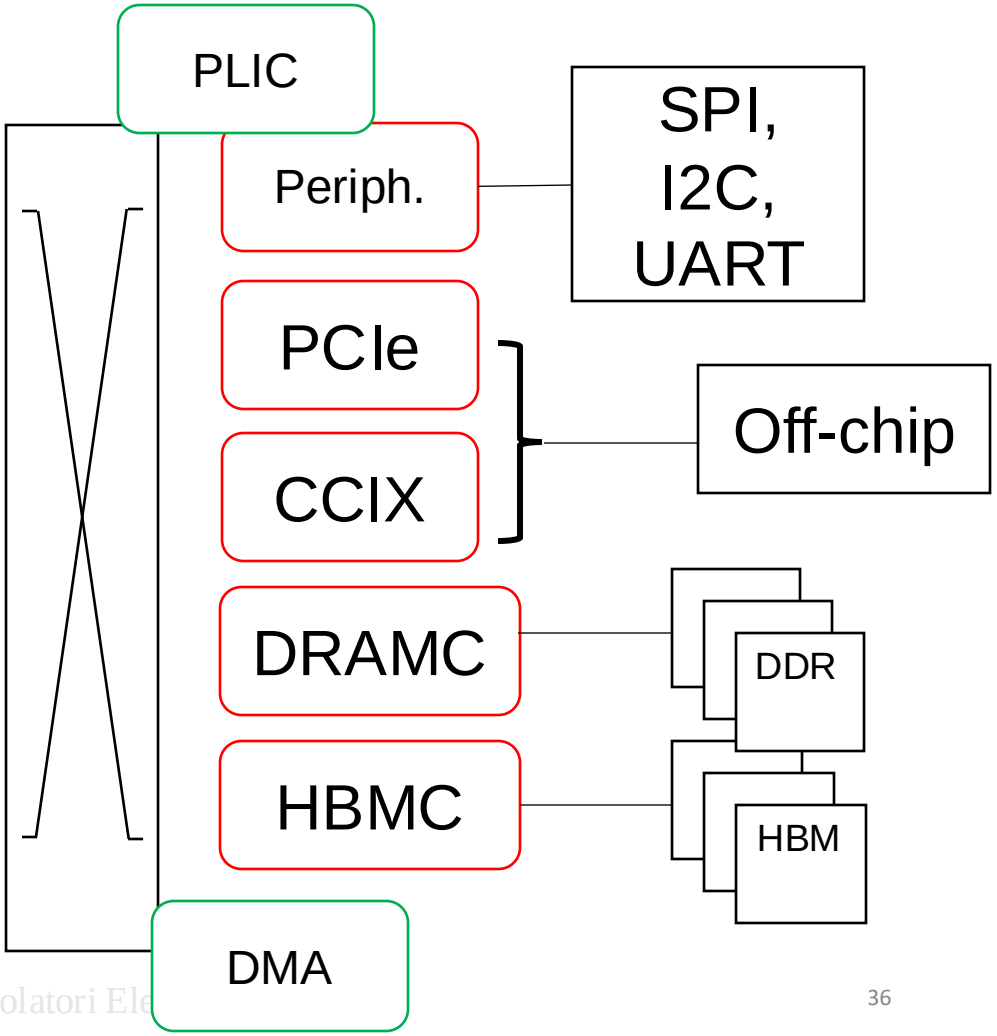


3. CPU ad alte prestazioni:
- ISA RISC-V
- Parallelismo a livello di istruzioni
- Parallelismo a livello di dati

Struttura di un calcolatore moderno

System on Chip Desing, many-core, SIMD/Vettoriale, integrazione 2.5D

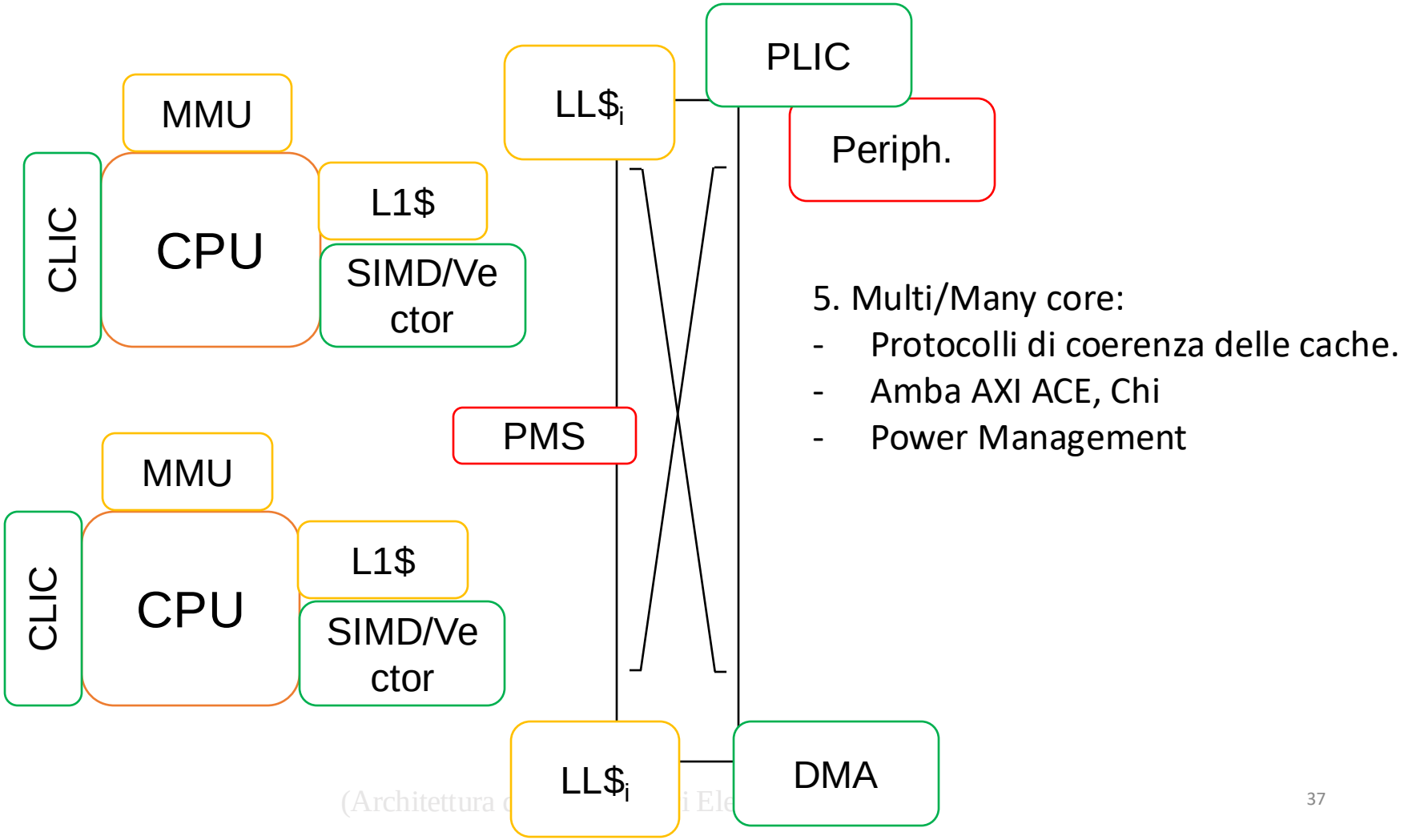




4. Periferiche on-chip/off-chip:
- Crossbar, Amba AXI bus
- PLIC/CLIC
- DMA

Struttura di un calcolatore moderno

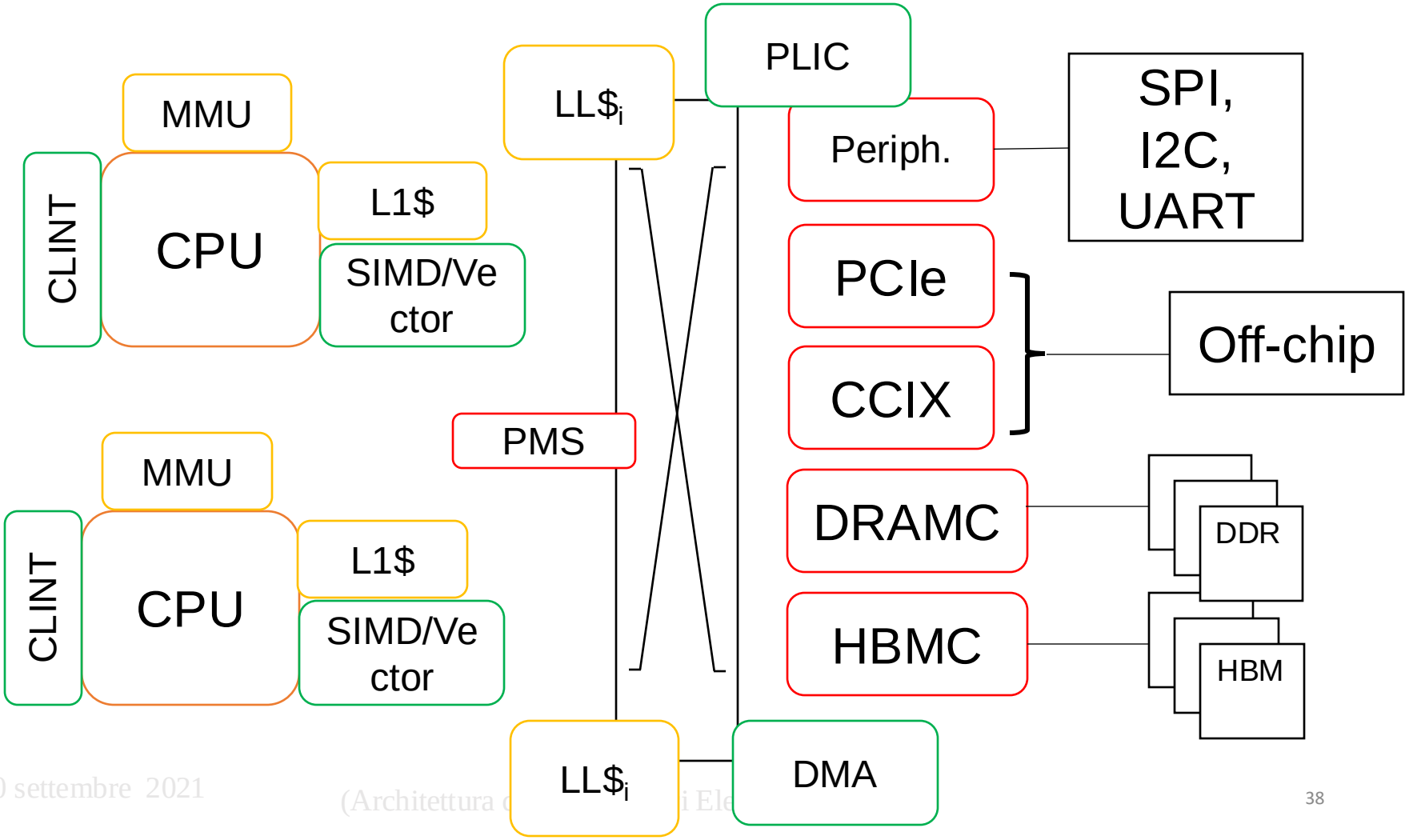
System on Chip Desing, many-core, SIMD/Vettoriale, integrazione 2.5D



5. Multi/Many core:
- Protocolli di coerenza delle cache.
- Amba AXI ACE, Chi
- Power Management

Struttura di un calcolatore moderno

System on Chip Desing, many-core, SIMD/Vettoriale, integrazione 2.5D



Modalità di svolgimento dell'esame
prova finale scritta

L'esame sarà costituito da una prova scritta di due ore ed una prova orale facoltativa.

La Prova Scritta vale 27/30esimi.

La Prova Orale vale 6/30esimi.

Lo scritto comprenderà 2 esercizi, uno sull'architettura di sistema e uno sull'architettura dell'unità di eleborazione.

La prova orale è facoltativa e vi ci possono accedere solo gli studenti che hanno superato lo scritto

Modalità di svolgimento dell'esame
prova finale scritta

La prova scritta comprenderà 2 esercizi, uno sull'architettura di sistema e uno sull'architettura dell'unità di eleborazione.

Modalità di svolgimento dell'esame
prova finale scritta

La prova orale è facoltativa e vi ci possono accedere solo gli studenti che hanno superato la prova scritta.

La prova orale consiste in un colloquio in cui lo studente dovrà valutare e discutere i principali vantaggi e svantaggi di di verse scelte architetturali discusse nel corso.

Il colloquio orale qualora concordato con il docente può essere sostituito dalla discussione di un elaborato su un approfondi mento di un argomento trattato nel corso o relativo all'attività progettuale qualora terminata. Il tema di eventuale approfondimento va concordato con il docente durante il corso.

Materiale didattico

- Il materiale utilizzato nel corso si trova sulla piattaforma IoL ed è così suddiviso:
  - **lucidi** presentati a lezione
  - **esercizi**
- Per approfondimenti si segnalano i seguenti testi:
  - Hennessy Patterson: "Computer architecture: a quantitative approach" - Morgan Kaufmann pub. Inc., Six edition (esiste anche la versione in italiano edita da Zanichelli)
  - David A Patterson John L Hennessy, Struttura e progetto dei calcolatori. Progettare con RISC-V
  - David J. Greaves, Modern System-or[illegible]1 Education Media

- Durante il corso verranno messi a disposizione link ad articoli, data sheet e altro materiale

David A Patterson, John L Hennessy

Struttura e progetto dei calcolatori

Progettare con RISC-V
Edizione italiana a cura di Alberto Borghese
2019

Ci occuperemo dunque
dell'Architettura e delle prestazioni
dell'unità di elaborazione (*aka* processore o CPU)

L'architettura della CPU è definita dalla seguente terna:

- Il set di istruzioni (architettura vista dall'utente, detta anche ISA (Instruction Set Architecture) o linguaggio macchina
- La struttura interna
- La realizzazione circuitale (cioè la tecnologia microelettronica impiegata nella realizzazione)

- Uno stesso set di istruzioni può essere realizzato con strutture interne diverse (es. x86-64, ARMv8, ARMv7, RISC-V)
- La stessa struttura interna può essere realizzata con tecnologie diverse (es. Haswell, Zen, Zen2, Zen3)
- **Fissato un benchmark (programma di riferimento), le prestazioni del calcolatore dipendono da tutte le componenti della terna**

$$CPU_{time} = N_{istruzioni} * CPI_{medio} * T_{ck}$$

ISA affermate sul mercato dei calcolatori

Nell'ambito di questo insegnamento considereremo solo le ISA oggi più affermate sul mercato dei calcolatori:

- ISA REGISTER REGISTER (R-R)
  - Esempio: RISC, ARM, RISC-V

Questo livello dell'architettura dei calcolatori è stato oggetto del corso di Calcolatori della laurea triennale

Strutture considerate nel corso

- Verranno considerate:
  - La struttura del principale componente della macchina di Von Neumann, e cioè l'Unità di Elaborazione (CPU)
  - la struttura del calcolatore, intesa come realizzazione del modello di riferimento (macchina di Von Neumann)

Corrispondenza architettura – velocità di esecuzione