NHR FAU

Friedrich-Alexander-Universität Erlangen-Nürnberg

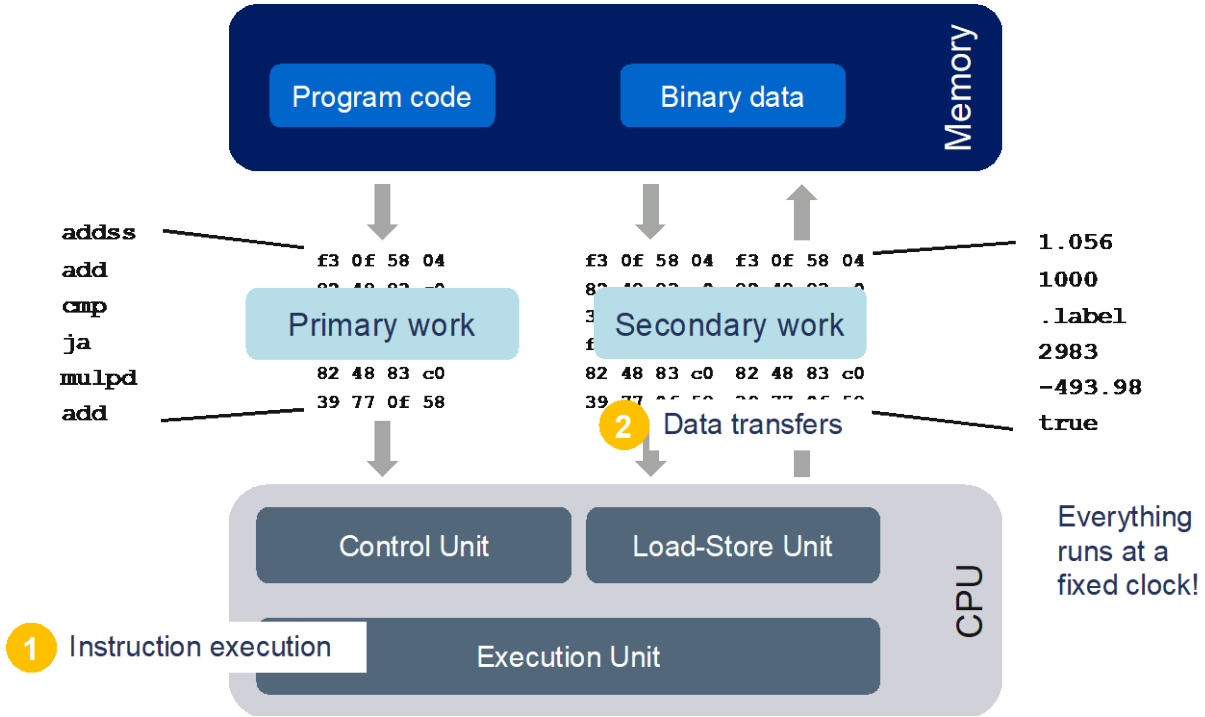
A review of processor advances over the past thirty years – an outsider's perspective

Jan Eitzinger

NHR PerfLab Seminar

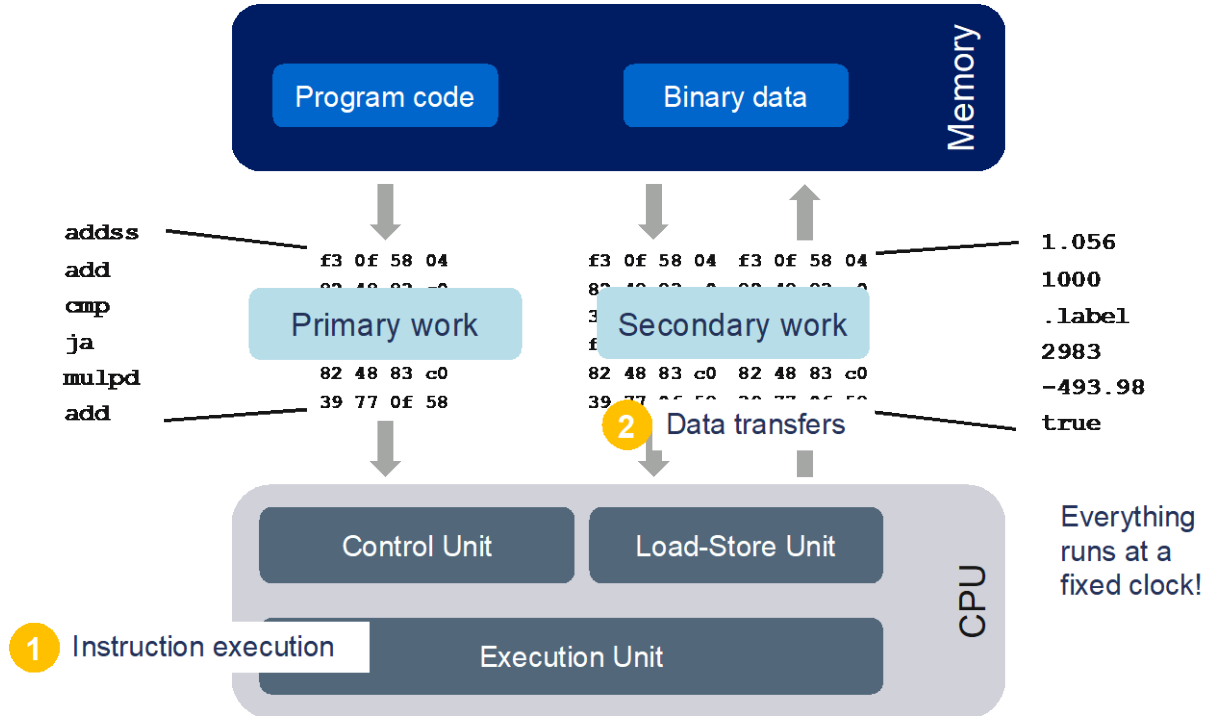
November 7, 2023

A bird's eye view



Stored Program Computer

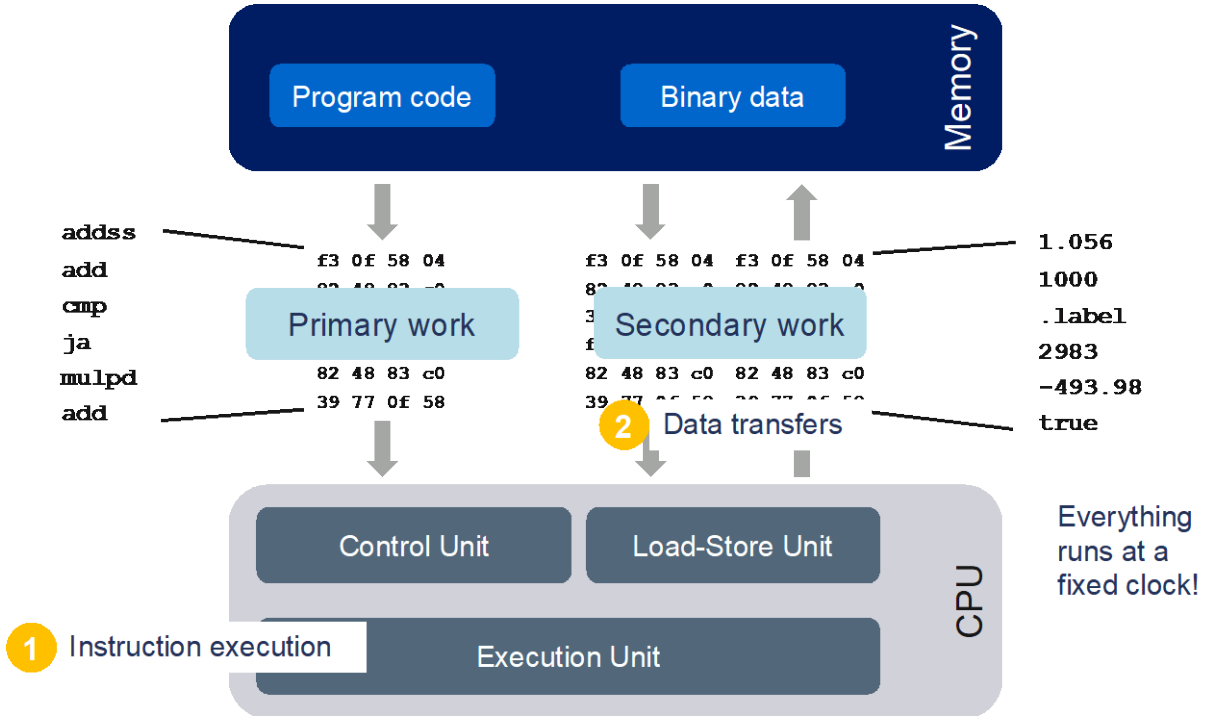
A bird's eye view



Focus on **relevant** software
Technical opportunities
Economical concerns
Marketing concerns

Stored Program Computer

A bird's eye view



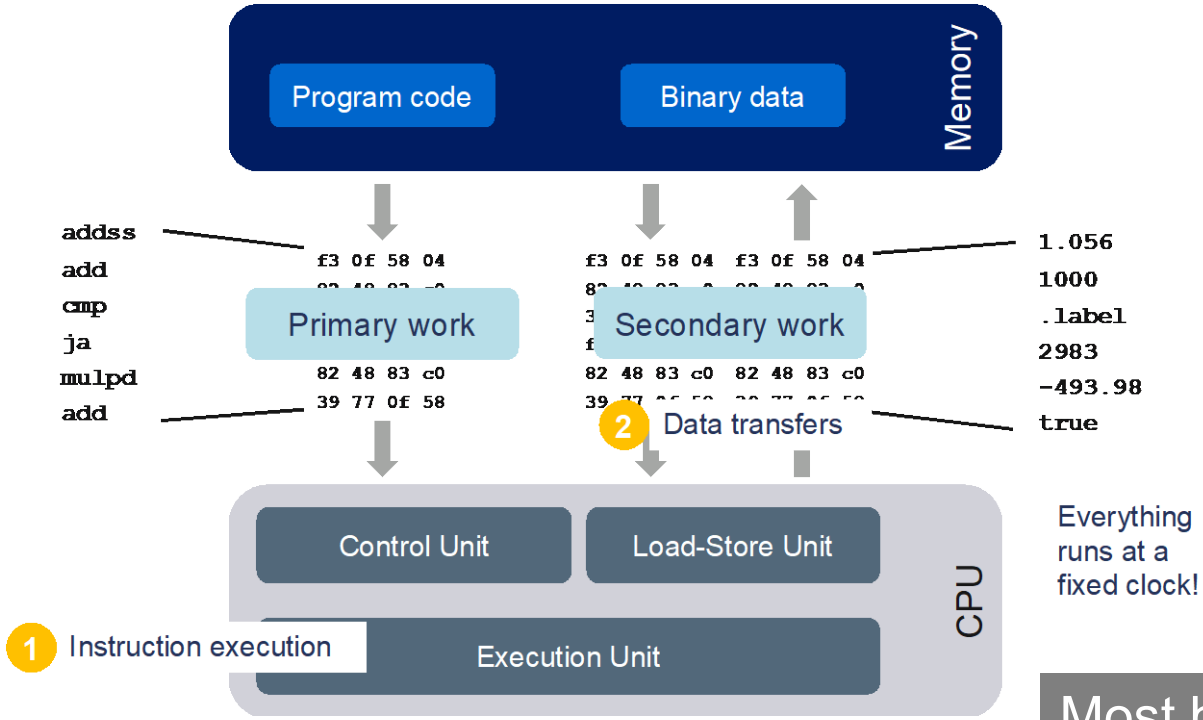
Stored Program Computer

Focus on **relevant** software
Technical opportunities
Economical concerns
Marketing concerns

Improvement strategies

- Increase clock speed
- Parallelism
- Specialization

A bird's eye view



Everything runs at a fixed clock!

Stored Program Computer

Focus on **relevant** software
Technical opportunities
Economical concerns
Marketing concerns

Improvement strategies

- Increase clock speed
- Parallelism
- Specialization

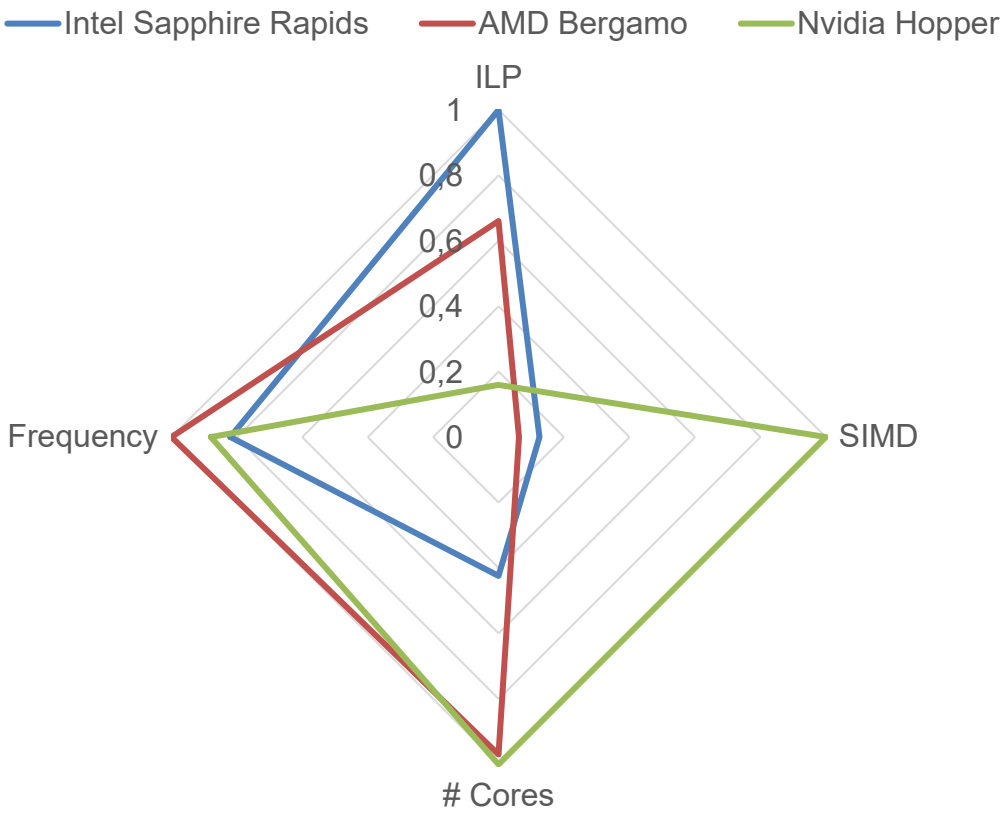
Most hardware optimizations put assumptions against the software!

Hardware-Software Co-Design?

- Hardware architects often experience the software as the messy side and tend to treat it as a blackbox
- This is also enforced by the fact that CPUs are later reviewed based on a small collection of given benchmarks
- For software developers hardware is often this complex, intransparent machinery that you have to live with

- This makes it very difficult to introduce innovative new solutions requiring both sides to change

Compromise within given power envelope



The run for higher frequencies or the *memory wall* (1990 – 2005)

- Enabled by advances in ILP technology and manufacturing processes

The run for higher frequencies or the *memory wall* (1990 – 2005)

- Enabled by advances in ILP technology and manufacturing processes

1990 x486
33 MHz
1000nm

The run for higher frequencies or the *memory wall* (1990 – 2005)

- Enabled by advances in ILP technology and manufacturing processes

1990 x486
33 MHz
1000nm

1996 P200
200 MHz
350nm 15W

The run for higher frequencies or the *memory wall* (1990 – 2005)

- Enabled by advances in ILP technology and manufacturing processes

1990 x486
33 MHz
1000nm

1996 P200
200 MHz
350nm 15W

1999 PIII
1.1 GHz
180nm 33W

The run for higher frequencies or the *memory wall* (1990 – 2005)

- Enabled by advances in ILP technology and manufacturing processes

1990 x486
33 MHz
1000nm

1996 P200
200 MHz
350nm 15W

1999 PIII
1.1 GHz
180nm 33W

2002 P4
2.5 GHz
130nm 61W

The run for higher frequencies or the *memory wall* (1990 – 2005)

- Enabled by advances in ILP technology and manufacturing processes

1990 x486
33 MHz
1000nm

1996 P200
200 MHz
350nm 15W

1999 PIII
1.1 GHz
180nm 33W

2002 P4
2.5 GHz
130nm 61W

2005 P4
3.8 GHz
90nm 115W

The run for higher frequencies or the *memory wall* (1990 – 2005)

- Enabled by advances in ILP technology and manufacturing processes

1990 x486
33 MHz
1000nm

1996 P200
200 MHz
350nm 15W

1999 PIII
1.1 GHz
180nm 33W

Factor 115 in 15 years!

2002 P4
2.5 GHz
130nm 61W

2005 P4
3.8 GHz
90nm 115W

The run for higher frequencies or the *memory wall* (1990 – 2005)

- Enabled by advances in ILP technology and manufacturing processes

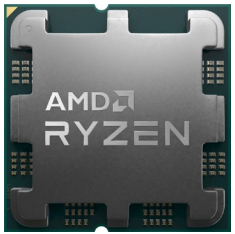
1990 x486
33 MHz
1000nm

1996 P200
200 MHz
350nm 15W

1999 PIII
1.1 GHz
180nm 33W

Factor 115 in 15 years!

2002 P4
2.5 GHz
130nm 61W



2023 AMD Ryzen
4.5 GHz
5nm 170W 16c

2005 P4
3.8 GHz
90nm 115W

The run for higher frequencies or the *memory wall* (1990 – 2005)

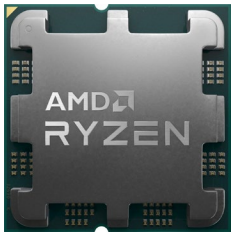
- Enabled by advances in ILP technology and manufacturing processes

1990 x486
33 MHz
1000nm

1996 P200
200 MHz
350nm 15W

1999 PIII
1.1 GHz
180nm 33W

Factor 115 in 15 years!



2023 AMD Ryzen
4.5 GHz
5nm 170W 16c

2002 P4
2.5 GHz
130nm 61W

2005 P4
3.8 GHz
90nm 115W

Intel Pentium 4 introduced DP FP SSE2 (2001) and Hyperthreading (2002). To profit you had to use SIMD instructions and parallelize your code!

SIMD or *free lunch is over* (1999 – 2018)

- Initially introduced in RISC workstations in early ninties

- Mainstream implementations targeted at multimedia and graphics
- 1997: MMX (64bit, Integer only)
- 1998: 3DNow! (64bit, SP FP)
- 1999: SSE (64bit, SP FP)
- 1999: Altivec/VMX (128bit, SP FP)
- 2001: SSE2 (128bit, DP FP)

Single-core DP floating-point performance

$$P_{core} = n_{super}^{FP} \cdot n_{FMA} \cdot n_{SIMD} \cdot f$$

- n_{super}^{FP}: Super-scalarity
- n_{FMA}: FMA factor
- n_{SIMD}: SIMD factor
- f: Clock Speed

Typical representatives	n_{super}^{FP} [inst./cy]	n_{FMA}	n_{SIMD} [ops/inst.]	@market	Ex. model	f [Gcy/s]	P_{core}[GF/s]
Nehalem	2	1	2	Q1/2009	X5570	2.93	11.7
Sandy Bridge	2	1	**4**	Q1/2012	E5-2680	2.7	21.6
Haswell	2	2	4	Q3/2014	E5-2695 v3	2.3	36.8
Skylake	2	2	**8**	Q3/2017	Gold 6148	2.0	64
AMD Zen	2	2	2	Q1/2017	Epyc 7451	2.3	18.4
AMD Zen2	2	2	4	Q4/2019	Epyc 7642	2.3	36.8
Fujitsu A64FX	2	2	8	Q2/2020	FX700	1.8	57.6
IBM POWER10	8	2	2	Q3/2020	-	3.5	112 (?)
Apple Silicon	4	2	2	Q1/2023	M2	3.5	56

Single-core DP floating-point performance

$$P_{core} = n_{super}^{FP} \cdot n_{FMA} \cdot n_{SIMD} \cdot f$$

Super-scalarity | FMA factor | SIMD factor | Clock Speed

Typical representatives	n_{super}^{FP} [inst./cy]	n_{FMA}	n_{SIMD} [ops/inst.]	@market	Ex. model	f [Gcy/s]	P_{core}[GF/s]
Nehalem	2	1	2	Q1/2009	X5570	2.93	11.7
Sandy Bridge	2	1	4	Q1/2012	E5-2680	2.7	21.6
Haswell	2	2	4	Q3/2014	E5-2695 v3	2.3	36.8
Skylake	2	2	8	Q3/2017	Gold 6148	2.0	64
AMD Zen	2	2	2	Q1/2017	Epyc 7451	2.3	18.4
AMD Zen2	2	2	4	Q4/2019	Epyc 7642	2.3	36.8
Fujitsu A64FX	2	2	8	Q2/2020	FX700	1.8	57.6
IBM POWER10	8	2	2	Q3/2020	-	3.5	112 (?)
Apple Silicon	4	2	2	Q1/2023	M2	3.5	56

Single-core DP floating-point performance

$$P_{core} = n_{super}^{FP} \cdot n_{FMA} \cdot n_{SIMD} \cdot f$$

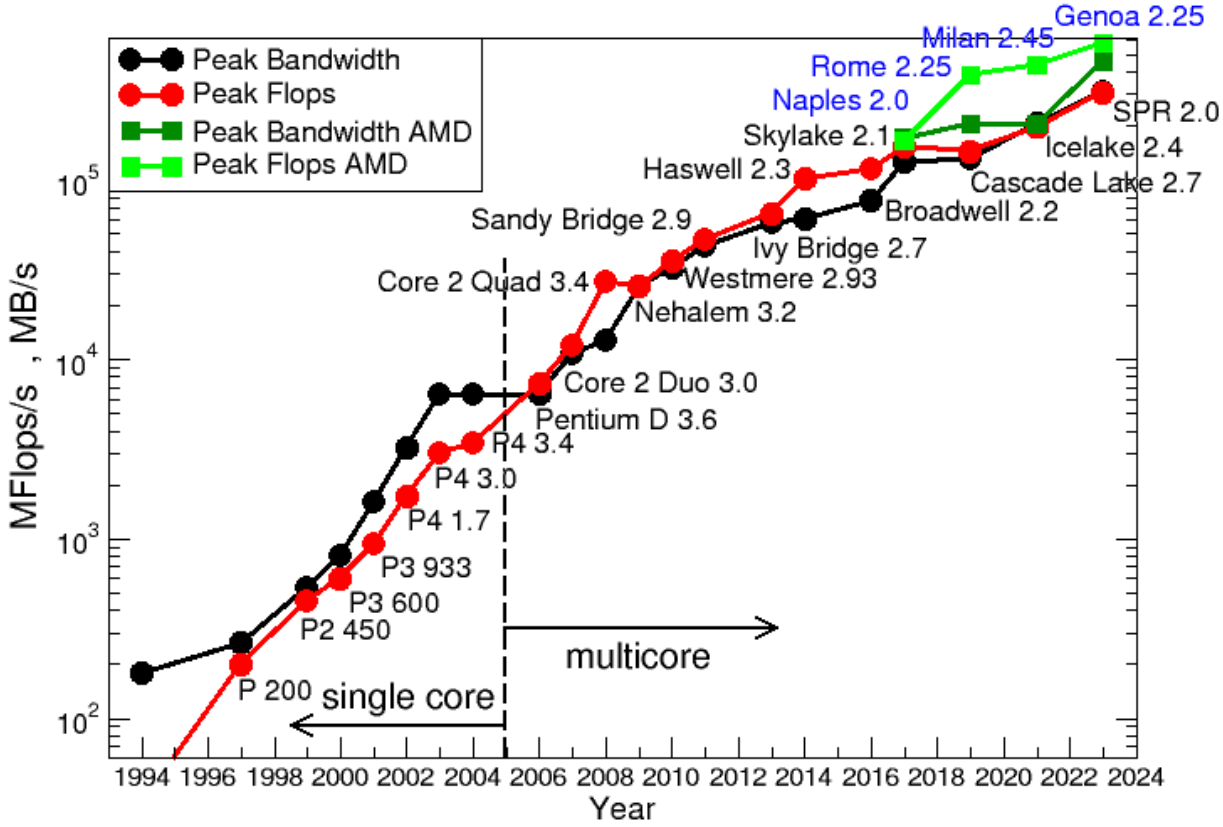
Super-scalarity | FMA factor | SIMD factor | Clock Speed

Typical representatives	n_{super}^{FP} [inst./cy]	n_{FMA}	n_{SIMD} [ops/inst.]	@market	Ex. model	f [Gcy/s]	P_{core}[GF/s]
Nehalem	2	1	2	Q1/2009	X5570	2.93	11.7
Sandy Bridge	2	1	4	Q1/2012	E5-2680	2.7	21.6
Haswell	2	2	4	Q3/2014	E5-2695 v3	2.3	36.8
Skylake	2	2	8	Q3/2017	Gold 6148	2.0	64
AMD Zen	2	2	2	Q1/2017	Epyc 7451	2.3	18.4
AMD Zen2	2	2	4	Q4/2019	Epyc 7642	2.3	36.8
Fujitsu A64FX	2	2	8	Q2/2020	FX700	1.8	57.6
IBM POWER10	8	2	2	Q3/2020	-	3.5	112 (?)
Apple Silicon	4	2	2	Q1/2023	M2	3.5	56

Multicore era *or the power wall* (2006 – 2022)

Challenge: Implement scalable interconnect!

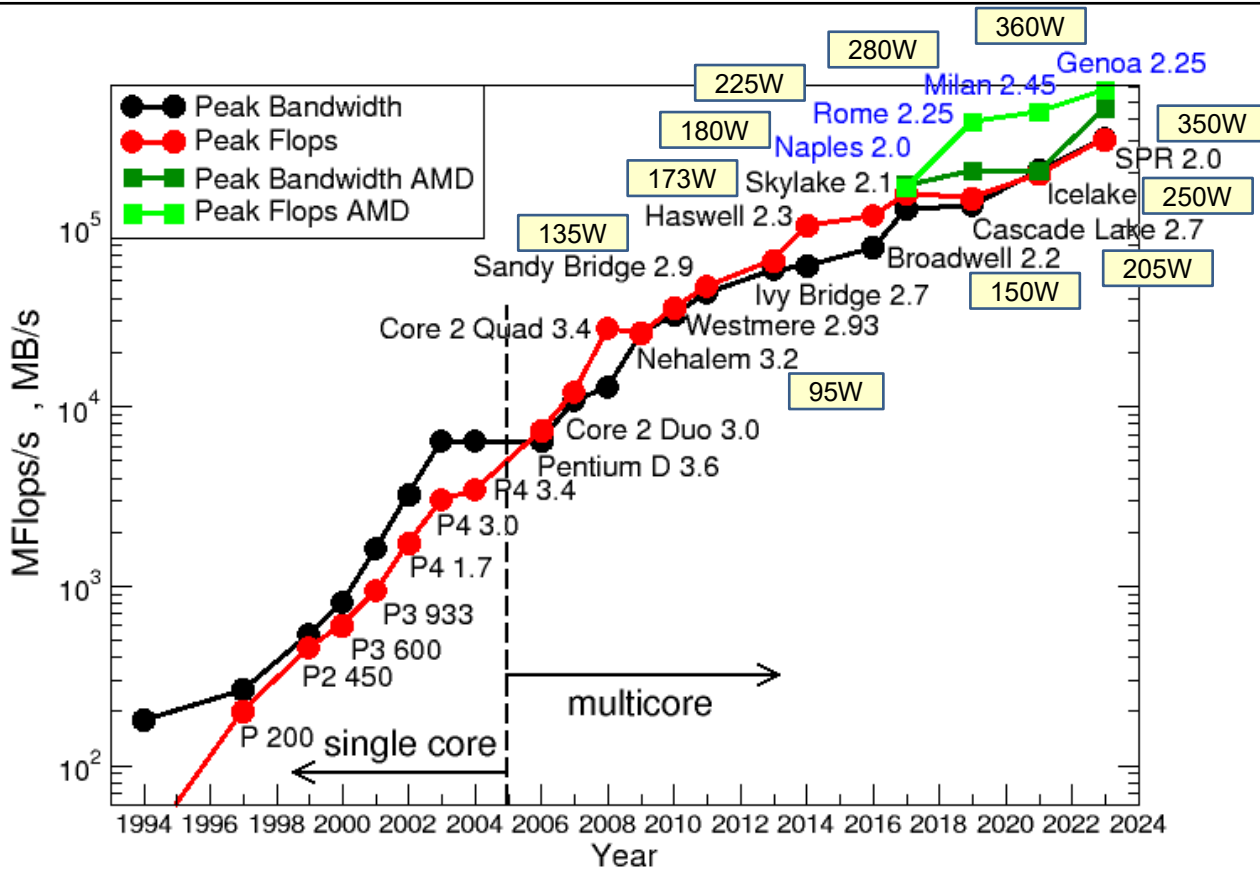
Iterations:



Multicore era *or the power wall* (2006 – 2022)

Challenge: Implement scalable interconnect!

Iterations:

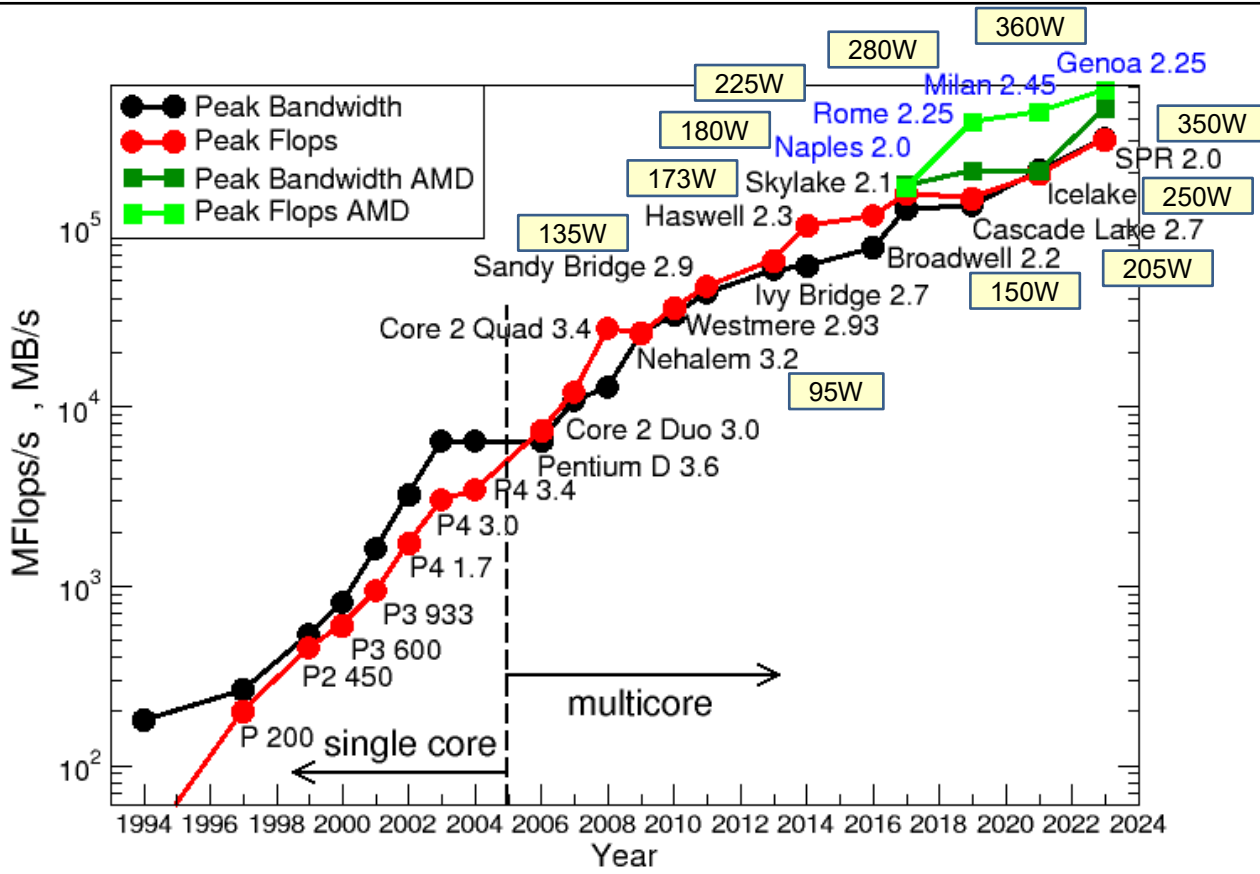


Multicore era *or the power wall* (2006 – 2022)

Challenge: Implement scalable interconnect!

Iterations:

- Direct connection

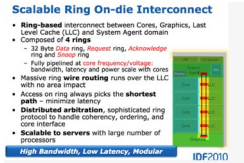


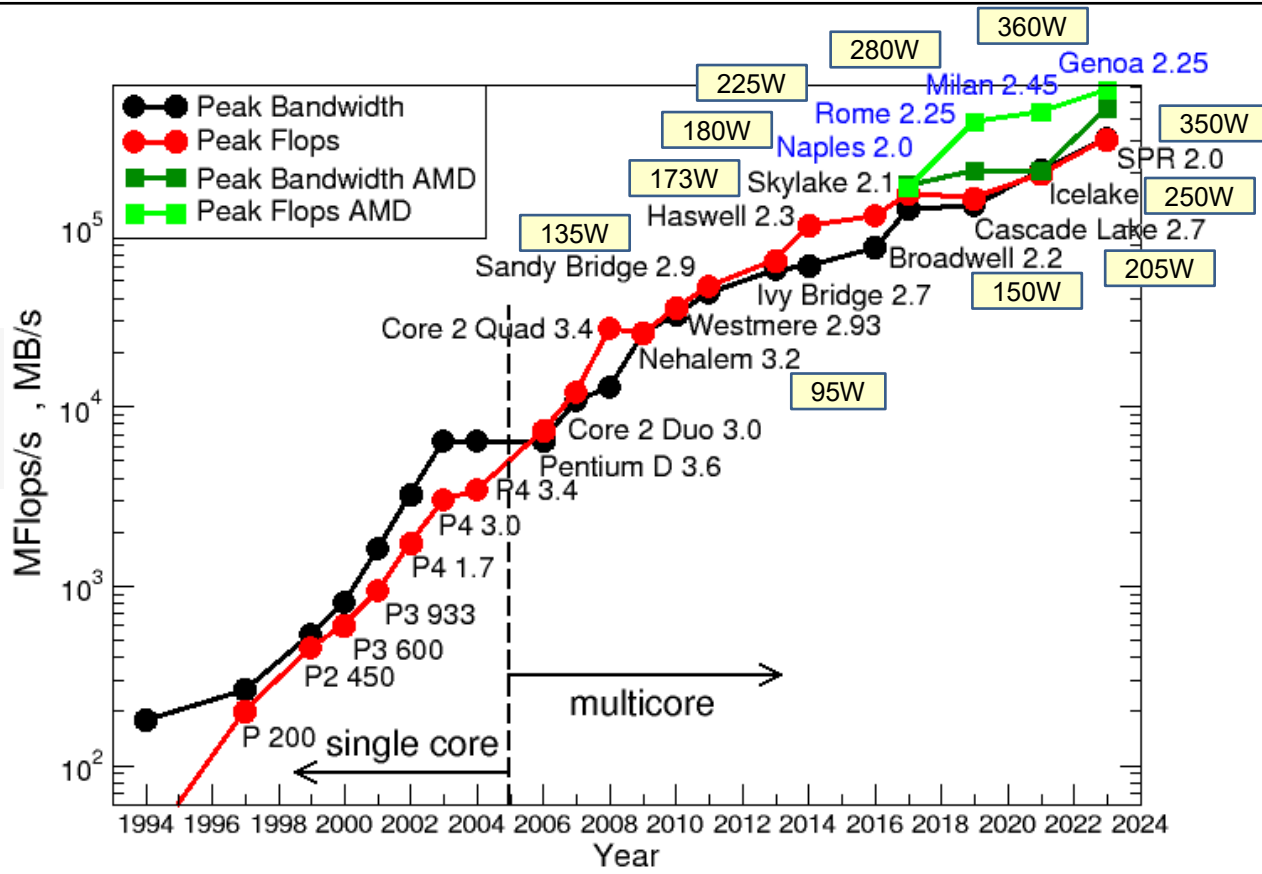
Multicore era *or the power wall* (2006 – 2022)

Challenge: Implement scalable interconnect!

Iterations:

- Direct connection
- Ring bus



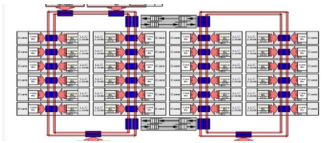


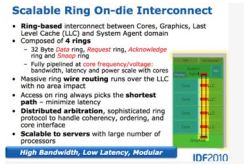
Multicore era *or the power wall* (2006 – 2022)

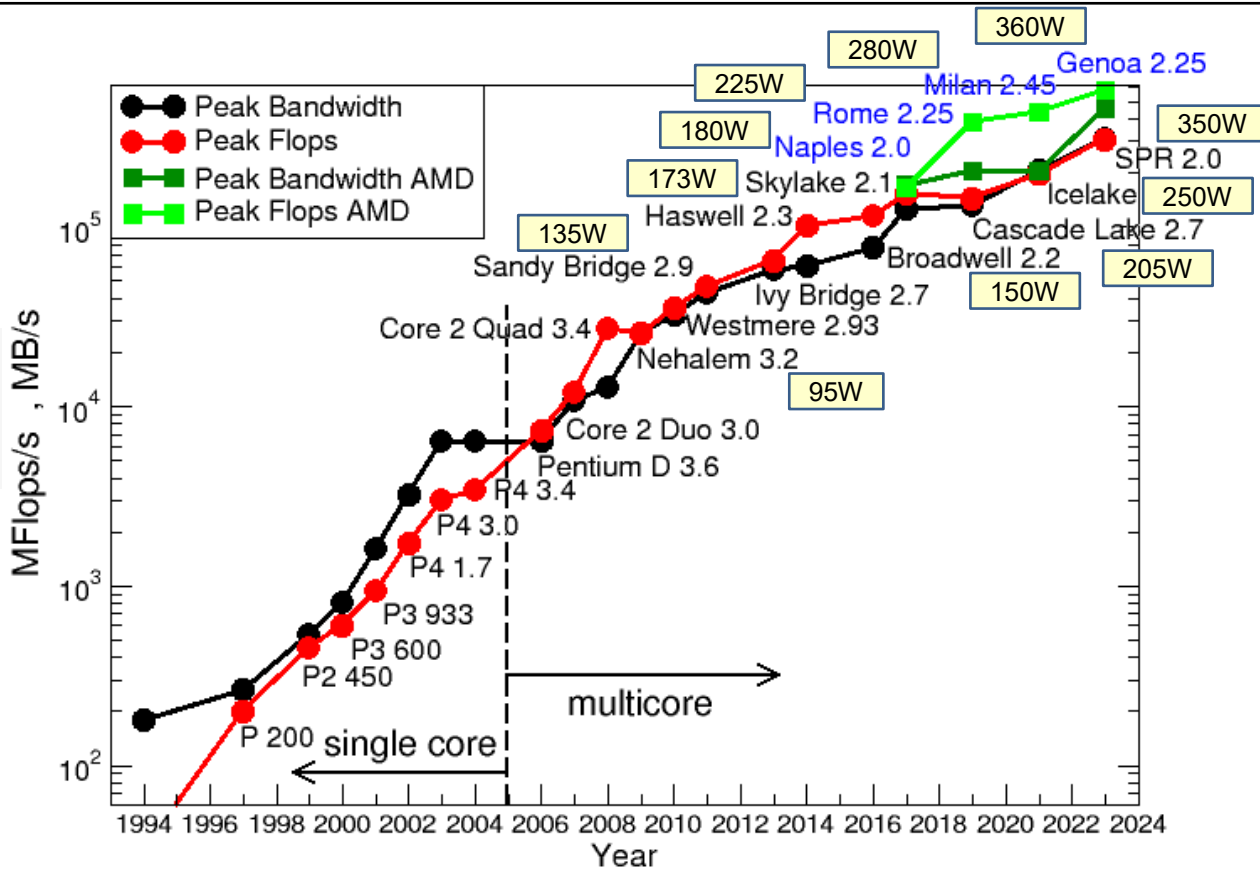
Challenge: Implement scalable interconnect!

Iterations:

- Direct connection
- Ring bus
- Two of them





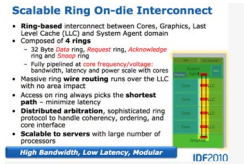


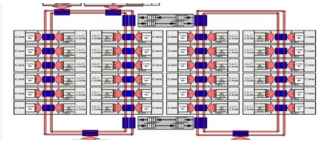
Multicore era *or the power wall* (2006 – 2022)

Challenge: Implement scalable interconnect!

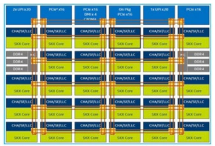
Iterations:

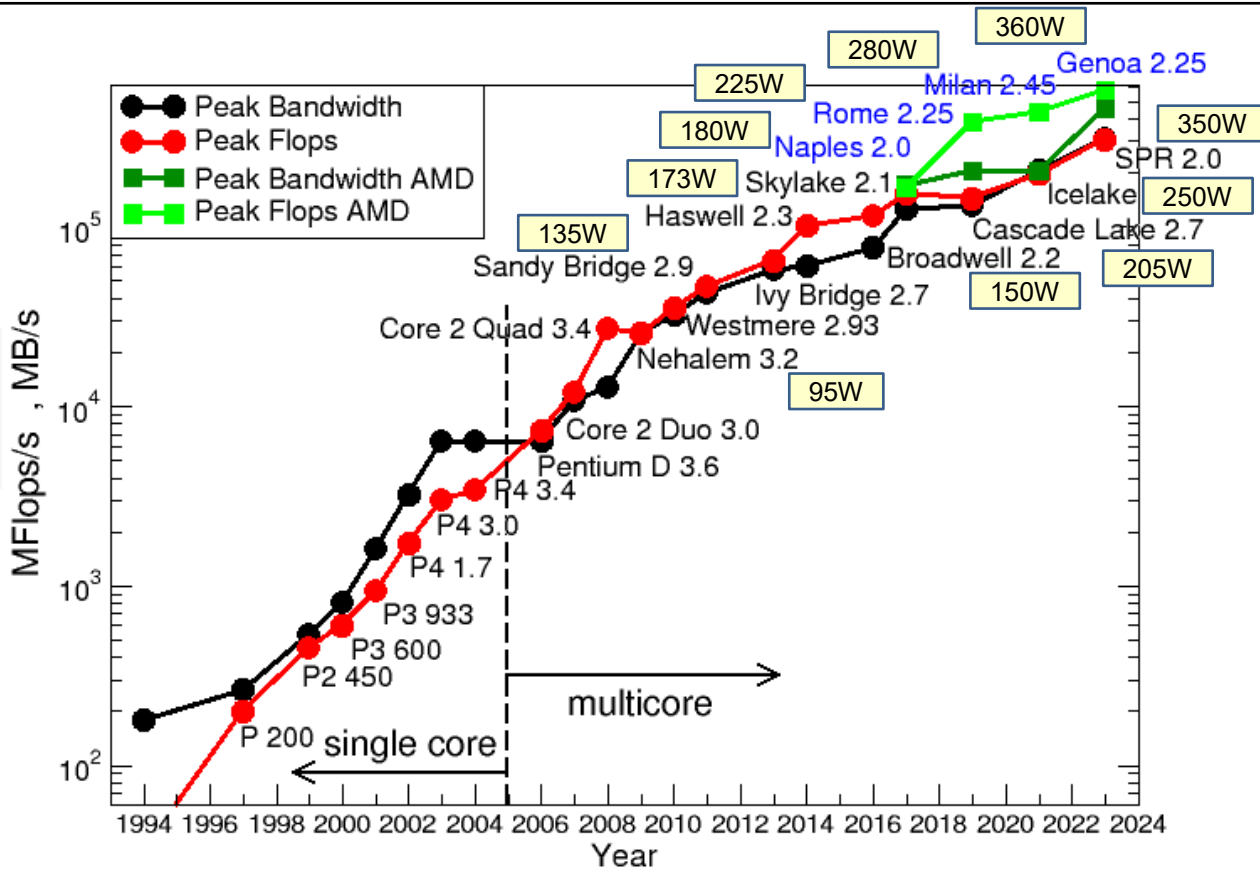
- Direct connection
- Ring bus
- Two of them





- Mesh



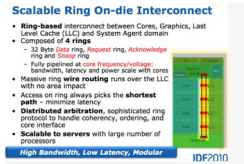


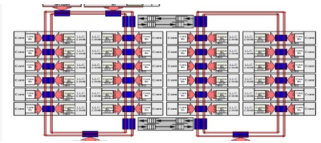
Multicore era *or the power wall* (2006 – 2022)

Challenge: Implement scalable interconnect!

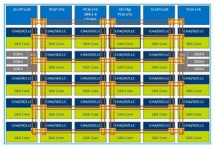
Iterations:

- Direct connection
- Ring bus
- Two of them

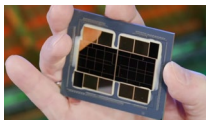


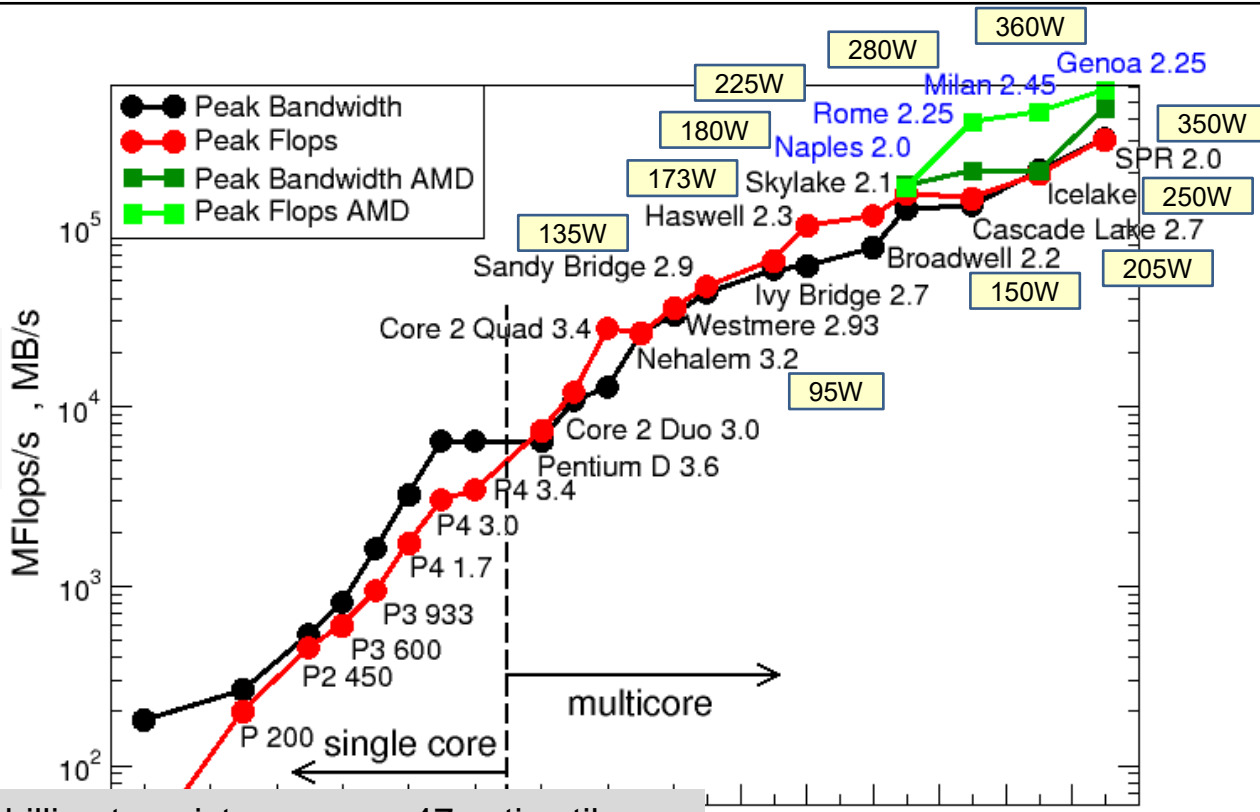


- Mesh



- Chiplets





100 billion transistors across 47 active tiles manufactured on five different process nodes

Chiplets save the day for multi-core

AMD Chiplets + IO Chip

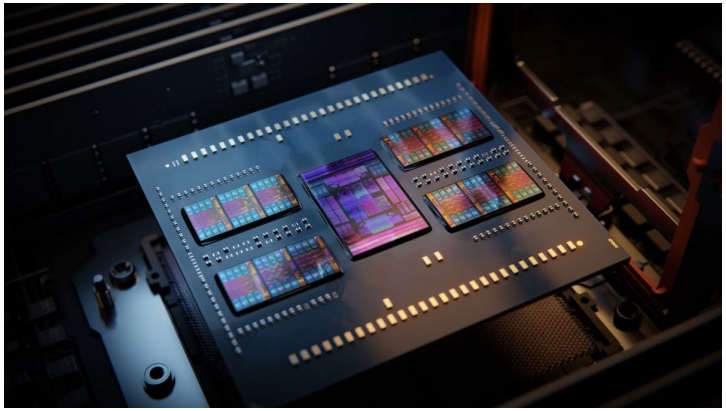
Naples 4 x 8c 14 nm

Rome 8 x 8c 7 nm

Milan 8 x 8c 7 nm

Genoa 12 x 8c

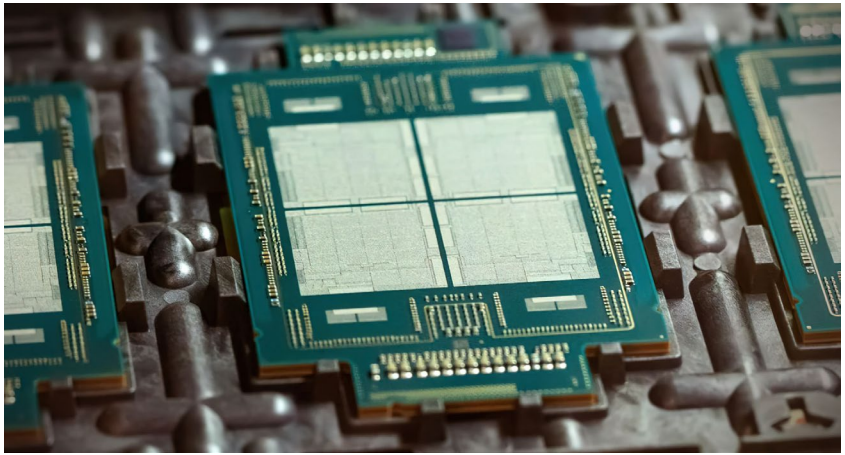
Bergamo 8 x 16c 5 nm 72 mm^2



Intel Sapphire Rapids

4 x 15c tiles Intel 7 400 mm^2

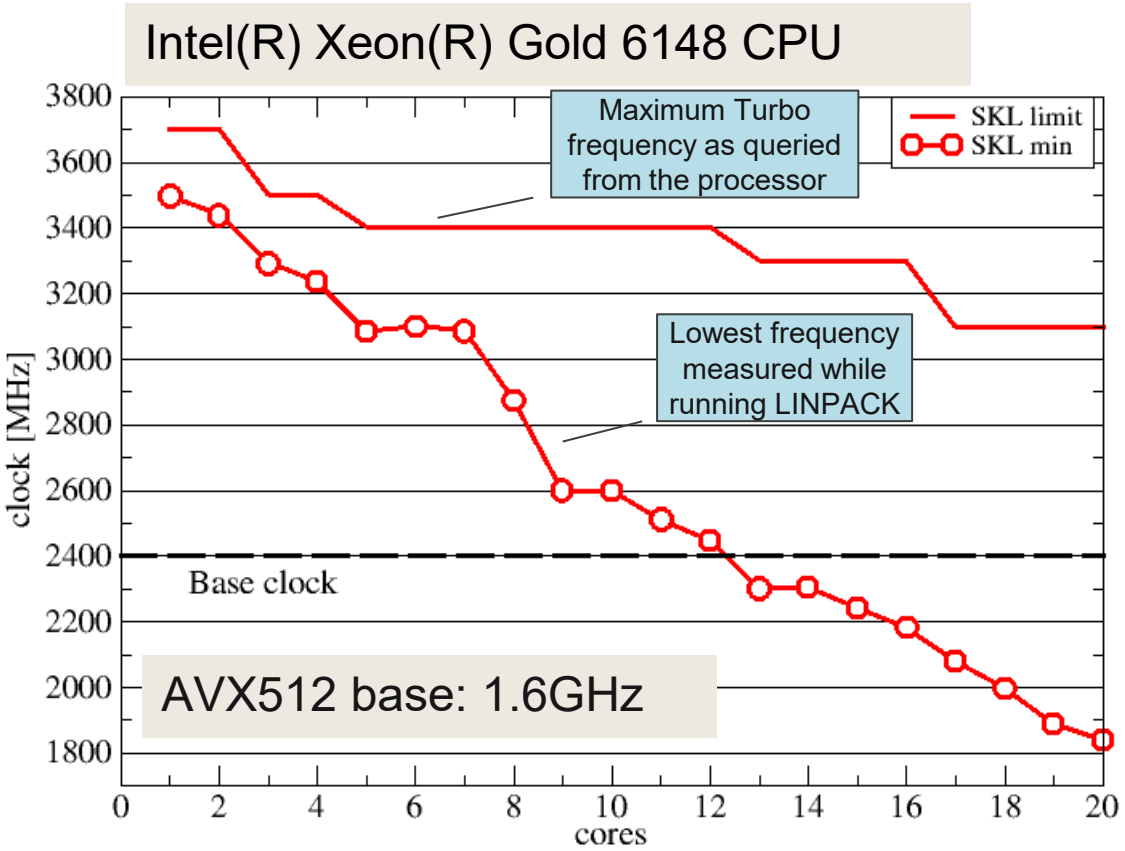
IO integrated in tiles



Which clock?

The processor dynamically overclocks to exploit more of the TDP envelope if fewer cores are active.

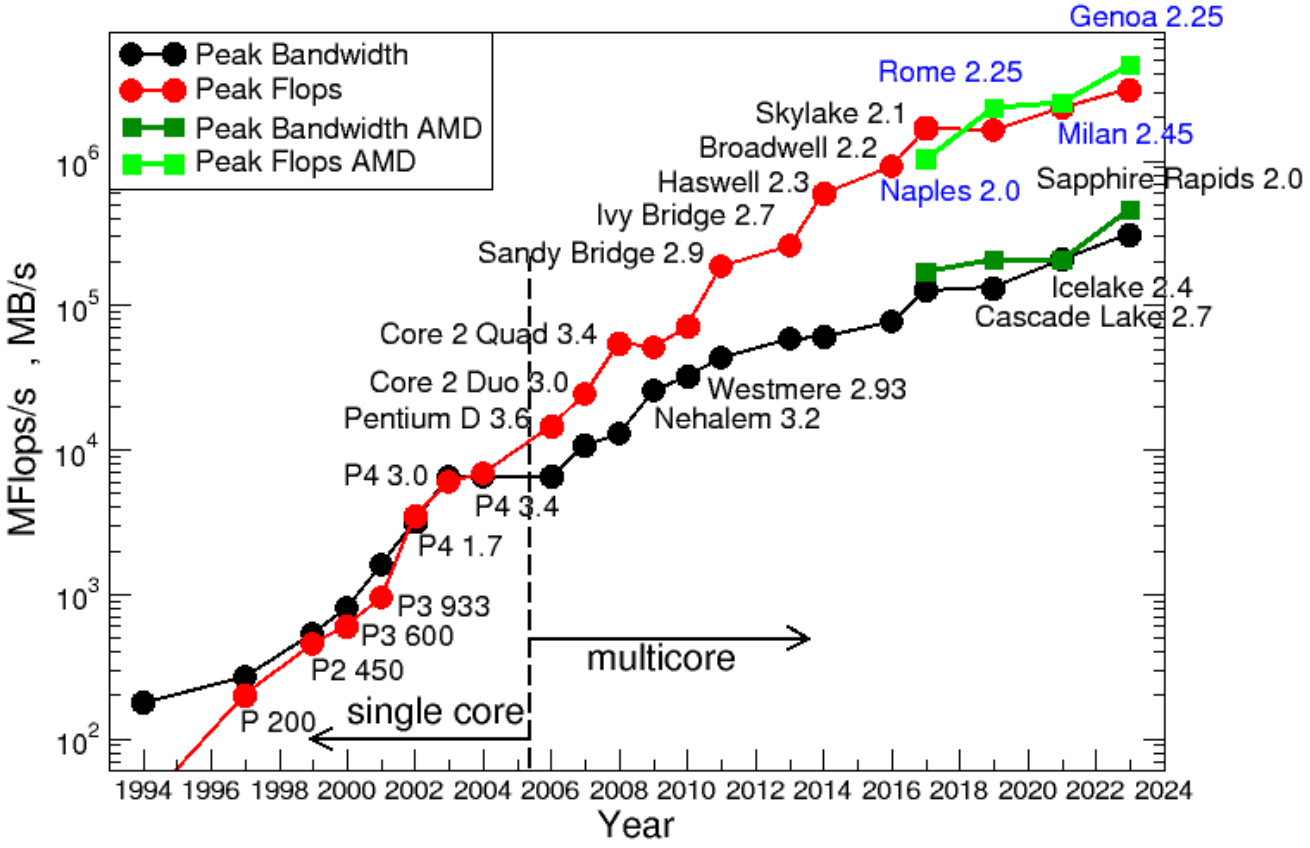
On Intel CPUs the base clock is meaningless!



The real picture

SIMD frequency measured using hot benchmark

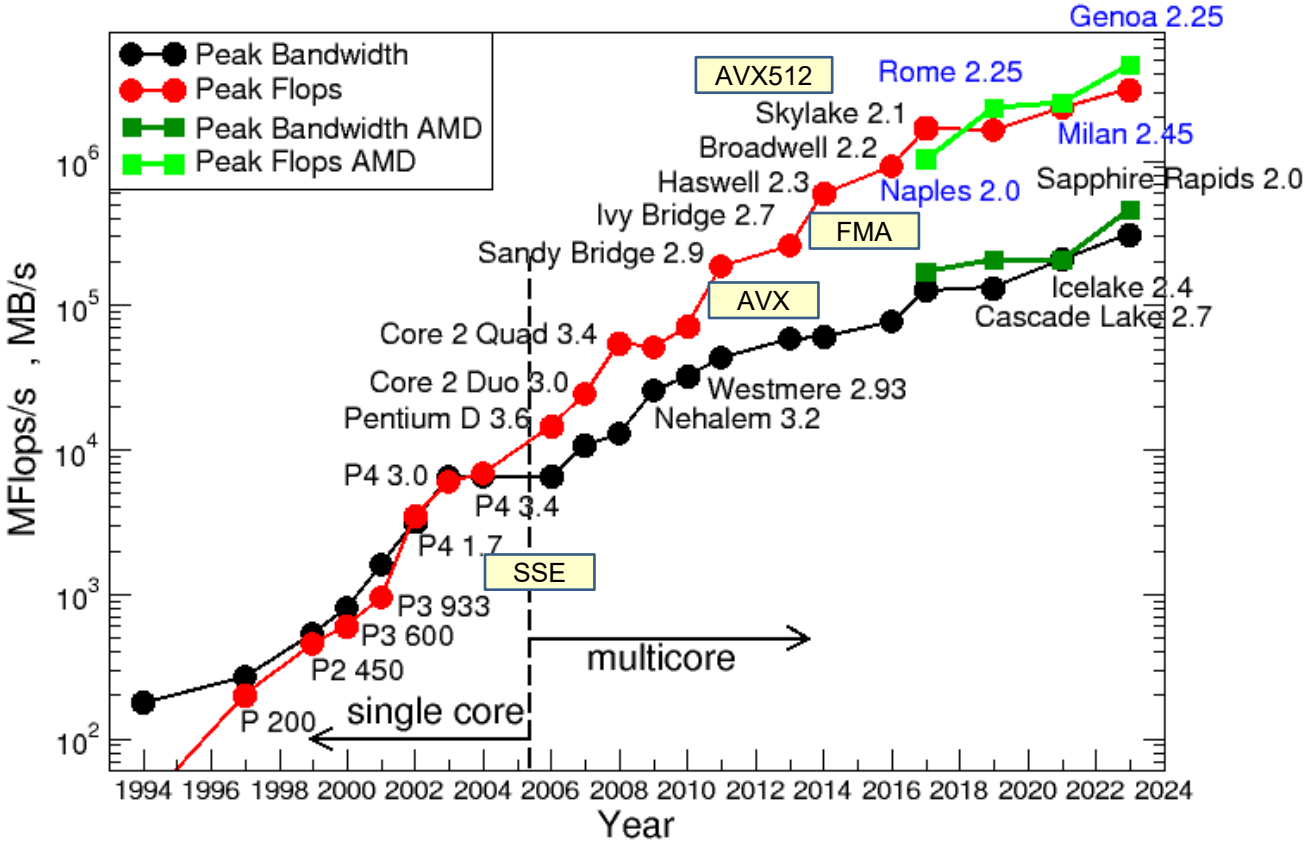
- SIMD enabled Intel to sustain performance increases!
- AMD is ahead in throughput and BW even without 64b SIMD



The real picture

SIMD frequency measured using hot benchmark

- SIMD enabled Intel to sustain performance increases!
- AMD is ahead in throughput and BW even without 64b SIMD



More of game with factors …

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	f_{simd}	Release	Model	P_{pkg} [GF/s]	TDP [W]	GF/ Watt
Sandy Bridge	2	1	4	8	2.9	2012	E5-2680	185	130	1.42
Haswell	2	2	4	18	2.0	2014	E5-2695-v3	576	120	4.80
Broadwell	2	2	4	22	2.2	2016	E5-2699-v4	774	145	5.34
Skylake	2	2	8	28	1.9	2017	8176	1702	165	10.31
Naples	2	2	4	32	2.0	2017		1024	180	5.69
Rome	2	2	4	64	2.25	2019		2304	225	10.24
Icelake	2	2	8	40	2.0	2021	8380	2560	270	9.48
SPR	2	2	8	56	1.9	2023	8470	3405	350	9.72
Bergamo	2	2	4	128	2.25	2023		4608	360	12.80
Apple Silicon	4	2	2	4+4	3.5	2023	M2	448	30	14.93

More of game with factors …

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	f_{simd}	Release	Model	P_{pkg} [GF/s]	TDP [W]	GF/ Watt
Sandy Bridge	2	1	4	8	2.9	2012	E5-2680	185	130	1.42
Haswell	2	2	4	18	2.0	2014	E5-2695-v3	576	120	4.80
Broadwell	2	2	4	22	2.2	2016	E5-2699-v4	774	145	5.34
Skylake	2	2	8	28	1.9	2017	8176	1702	165	10.31
Naples	2	2	4	32	2.0	2017		1024	180	5.69
Rome	2	2	4	64	2.25	2019		2304	225	10.24
Icelake	2	2	8	40	2.0	2021	8380	2560	270	9.48
SPR	2	2	8	56	1.9	2023	8470	3405	350	9.72
Bergamo	2	2	4	128	2.25	2023		4608	360	12.80
Apple Silicon	4	2	2	4+4	3.5	2023	M2	448	30	14.93

More of game with factors …

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	f_{simd}	Release	Model	P_{pkg} [GF/s]	TDP [W]	GF/ Watt
Sandy Bridge	2	1	4	8	2.9	2012	E5-2680	185	130	1.42
Haswell	2	2	4	18	2.0	2014	E5-2695-v3	576	120	4.80
Broadwell	2	2	4	22	2.2	2016	E5-2699-v4	774	145	5.34
Skylake	2	2	8	28	1.9	2017	8176	1702	165	10.31
Naples	2	2	4	32	2.0	2017		1024	180	5.69
Rome	2	2	4	64	2.25	2019		2304	225	10.24
Icelake	2	2	8	40	2.0	2021	8380	2560	270	9.48
SPR	2	2	8	56	1.9	2023	8470	3405	350	9.72
Bergamo	2	2	4	128	2.25	2023		4608	360	12.80
Apple Silicon	4	2	2	4+4	3.5	2023	M2	448	30	14.93

More of game with factors …

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	f_{simd}	Release	Model	P_{pkg} [GF/s]	TDP [W]	GF/ Watt
Sandy Bridge	2	1	4	8	2.9	2012	E5-2680	185	130	1.42
Haswell	2	2	4	18	2.0	2014	E5-2695-v3	576	120	4.80
Broadwell	2	2	4	22	2.2	2016	E5-2699-v4	774	145	5.34
Skylake	2	2	8	28	1.9	2017	8176	1702	165	10.31
Naples	2	2	4	32	2.0	2017		1024	180	5.69
Rome	2	2	4	64	2.25	2019		2304	225	10.24
Icelake	2	2	8	40	2.0	2021	8380	2560	270	9.48
SPR	2	2	8	56	1.9	2023	8470	3405	350	9.72
Bergamo	2	2	4	128	2.25	2023		4608	360	12.80
Apple Silicon	4	2	2	4+4	3.5	2023	M2	448	30	14.93

What about main memory?

Same game with factors:

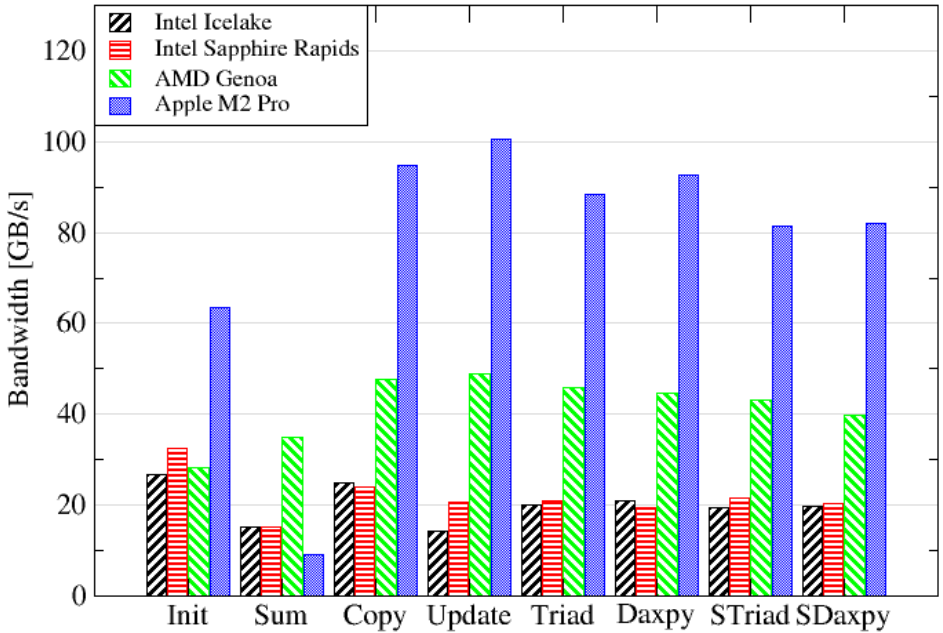
$$BW_{pkg} = n_{interfaces} \cdot n_{channels} \cdot 16 \cdot f_{bus}$$

AMD Genoa: 4 interfaces, 3 channels each, DDR5-4800 (2.4GHz)
460.8 GB/s

Intel Sapphire Rapids: 4 interfaces, 2 channels each, DDR5-4800 (2.4GHz)
307.2 GB/s

Apple M2 Ultra: 4 interfaces, 4 channels each, LPDDR5-6400 (3.2GHz)
819.2 GB/s

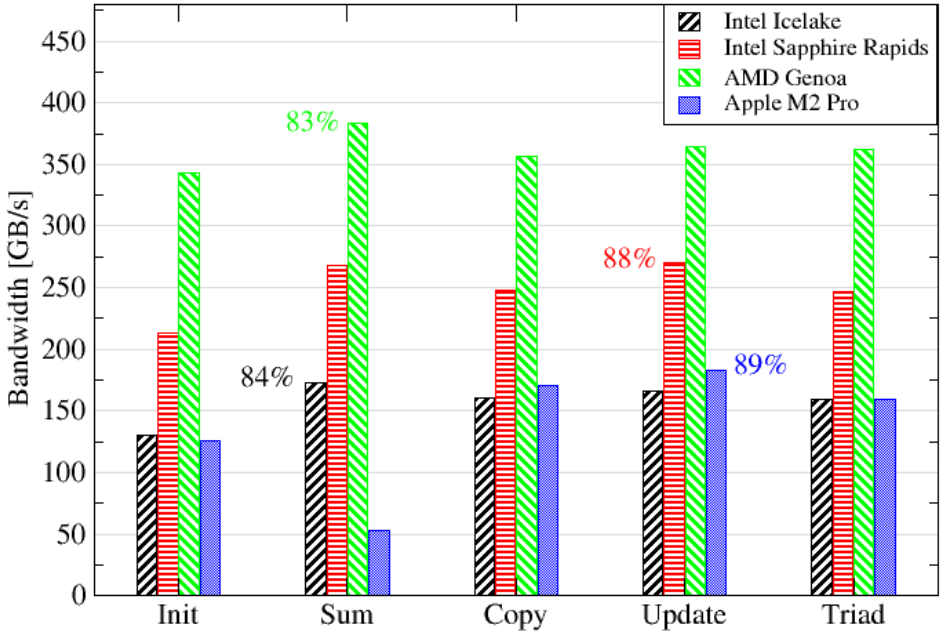
Single thread memory bandwidth



- Increasing number of load streams, with and without store miss.
- Intel architectures have history of low single thread bandwidth.
- AMD and Intel sequential bandwidth is a function of core architecture
- Apple achieves 50% of total bandwidth with single thread!

https://github.com/RRZE-HPC/TheBandwidthBenchmark

Package total memory bandwidth



- Total package bandwidth using all available cores
- AMD is far ahead of the competition (with DDR RAM)
- Apple bandwidth is on same level than previous generation Icelake

https://github.com/RRZE-HPC/TheBandwidthBenchmark

Accelerators or *the parallel wall* (2007 – ??)

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	f	Release	Model	P_{pkg} [GF/s]	TDP	GF/Watt
Sandy Bridge	2	1	4	8		2012	E5-2680	173	130	1,33
Pascal	1	2	32	56	1.480	2016	GP100	5304	300	17.68
Skylake	2	2	8	26	1.85	2017	8170	1581	165	9,58
Volta	1	2	32	80	1.597	2017	GV100	8177	300	27.25
Ampere	1	2	32	108	1.410	2020	GA100	9746	400	24.36
CDNA 1	1	2	32	120	1.50	2020	MI100	11520	300	38.40
CDNA 2	1	2	64	2x110	1.70	2021	MI250x	47872	560	85.48
Hopper	2	2	32	132	1.980	2022	GH100	33450	700	47.78
Bergamo	2	2	4	128	2.25	2023	9754	4608	360	12.80
Ponte Vecchio	8	2	16	128	1.60	2023	Max 1550	52430	600	87.38

Accelerators or *the parallel wall* (2007 – ??)

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	f	Release	Model	P_{pkg} [GF/s]	TDP	GF/Watt
Sandy Bridge	2	1	4	8		2012	E5-2680	173	130	1,33
Pascal	1	2	32	56	1.480	2016	GP100	5304	300	17.68
Skylake	2	2	8	26	1.85	2017	8170	1581	165	9,58
Volta	1	2	32	80	1.597	2017	GV100	8177	300	27.25
Ampere	1	2	32	108	1.410	2020	GA100	9746	400	24.36
CDNA 1	1	2	32	120	1.50	2020	MI100	11520	300	38.40
CDNA 2	1	2	64	2x110	1.70	2021	MI250x	47872	560	85.48
Hopper	2	2	32	132	1.980	2022	GH100	33450	700	47.78
Bergamo	2	2	4	128	2.25	2023	9754	4608	360	12.80
Ponte Vecchio	8	2	16	128	1.60	2023	Max 1550	52430	600	87.38

Accelerators or *the parallel wall* (2007 – ??)

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	f	Release	Model	P_{pkg} [GF/s]	TDP	GF/Watt
Sandy Bridge	2	1	4	8		2012	E5-2680	173	130	1,33
Pascal	1	2	32	56	1.480	2016	GP100	5304	300	17.68
Skylake	2	2	8	26	1.85	2017	8170	1581	165	9,58
Volta	1	2	32	80	1.597	2017	GV100	8177	300	27.25
Ampere	1	2	32	108	1.410	2020	GA100	9746	400	24.36
CDNA 1	1	2	32	120	1.50	2020	MI100	11520	300	38.40
CDNA 2	1	2	64	2x110	1.70	2021	MI250x	47872	560	85.48
Hopper	2	2	32	132	1.980	2022	GH100	33450	700	47.78
Bergamo	2	2	4	128	2.25	2023	9754	4608	360	12.80
Ponte Vecchio	8	2	16	128	1.60	2023	Max 1550	52430	600	87.38

Accelerators or *the parallel wall* (2007 – ??)

uArch	n_{super}^{FP}	n_{FMA}	n_{SIMD}	n_{cores}	f	Release	Model	P_{pkg} [GF/s]	TDP	GF/Watt
Sandy Bridge	2	1	4	8		2012	E5-2680	173	130	1,33
Pascal	1	2	32	56	1.480	2016	GP100	5304	300	17.68
Skylake	2	2	8	26	1.85	2017	8170	1581	165	9,58
Volta	1	2	32				GV100	8177	300	27.25
Ampere	1	2	32				GA100	9746	400	24.36
CDNA 1	1	2	32				MI100	11520	300	38.40
CDNA 2	1	2	64	2x110	1.70	2021	MI250x	47872	560	85.48
Hopper	2	2	32	132	1.980	2022	GH100	33450	700	47.78
Bergamo	2	2	4	128	2.25	2023	9754	4608	360	12.80
Ponte Vecchio	8	2	16	128	1.60	2023	Max 1550	52430	600	87.38

Double that in case you can use Tensor cores

Specialized execution engines

There already existed special purpose instructions for graphics and cryptography operations

Modern system-on-chip designs add execution engines

- Apple Neural Engine and Media Engine for video transcoding
- Nvidia Tensor cores
- Power 10 MMA engine

There are also standalone chips for AI, autonomous driving and crypto

- Google TPU
- ARM MLP
- Tesla FSD

A view on the Apple silicon processor architecture

- Apple started in 2012 to use their own chip design in their mobile devices
- Complete stack under control allows for more freedom in chip design

- This results in unique design decisions
  - Focus on ILP and extreme level of superscalar execution
  - Very capable memory hierarchy with large low latency caches and memory bandwith unseen in desktop chips before
  - System-on-chip design with very capable integrated GPU and special purpose hardware for ML and Video transcoding

- Same design since 2020 also used in Apple desktop devices

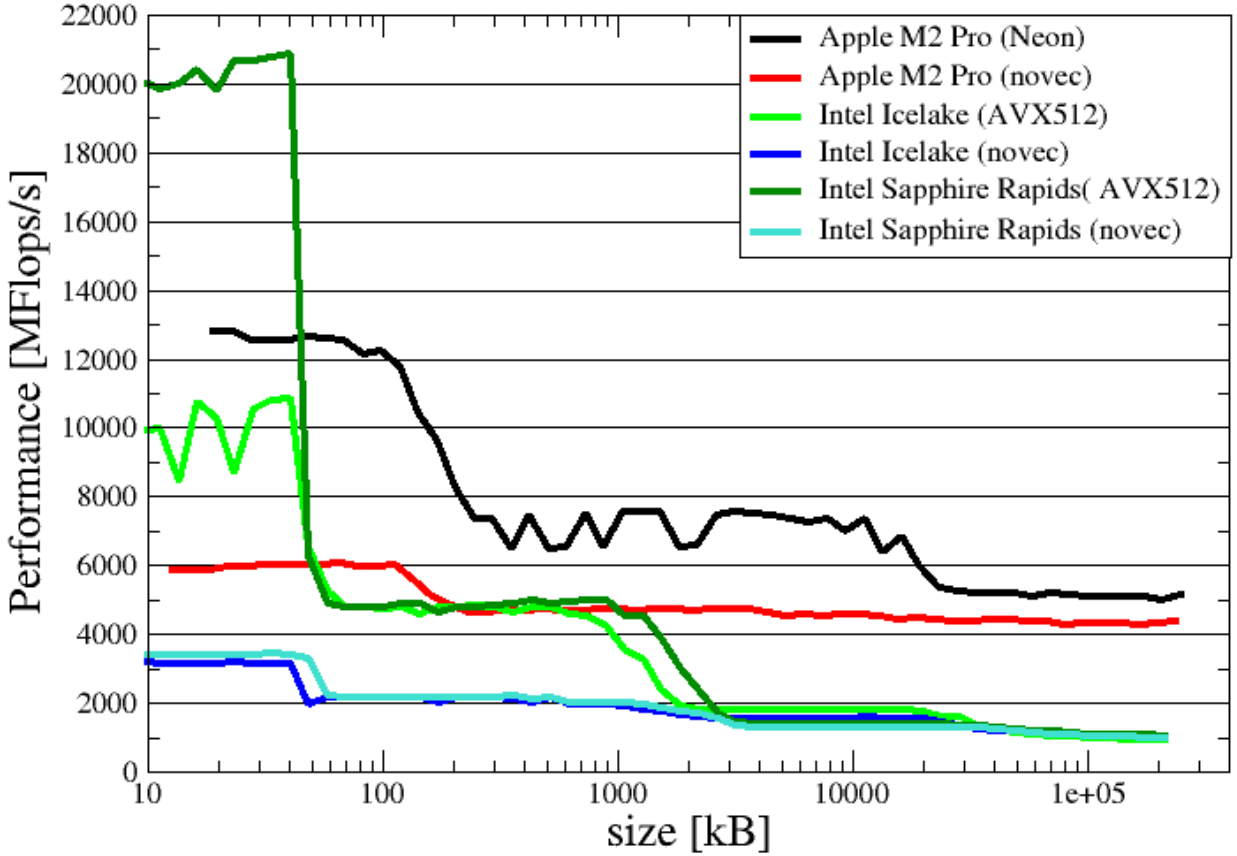
Comparison vs Intel Sapphire Rapids

Property	Intel Sapphire Rapids	Apple Silicon M2
Core count	56	Up to 12 (8P + 4E) Max
Frequency	2.0 GHz (Turbo 3.8 GHz)	3.5 GHz (700MHz – 3.5GHz)
L1D / I size	48 kB / 32 kB (5cy)	128 kB / 196 kB (3cy)
L2 size	2 MB (17cy)	32 MB (17cy, shared) + 4MB (shared)
L3 size	105 MB (50+cy, shared)	-
Memory	270 GB/s (25 GB/s) DDR5@4800 8-Channel	182 GB/s (101.5 GB/s) LPDDR5@6400 4-Channel
TDP	350W	30W
Superscalarity	6 (2 FP)	8 (4 FP)
SIMD width	AVX512 64b	Neon 16b
Load + Store	2L + 1S	3L+1S / 2L + 2S
FMA units	2	4

Schoenauer triad: Sequential data bandwidth

Data access pattern
3L + 1S per iteration

- M2 novec L2 bw is on same level as Intel AVX512 case
- M2 shows very high performance with no drop to main memory
- Intel LLC no target cache for HPC



Application benchmarks, single core

MD-Bench, Verlet list solver, Lennard-Jones potential

Processor	Overall / Force kernel / Neighbour list
Icelake	5.64s / 2.50s / 2.95s 4.65 MAups
Sapphire Rapids	5.29s / 2.30s / 2.84s 4.95 MAups
M2 Pro	5.17s / 2.51s / **2.32s** 5.07 MAups

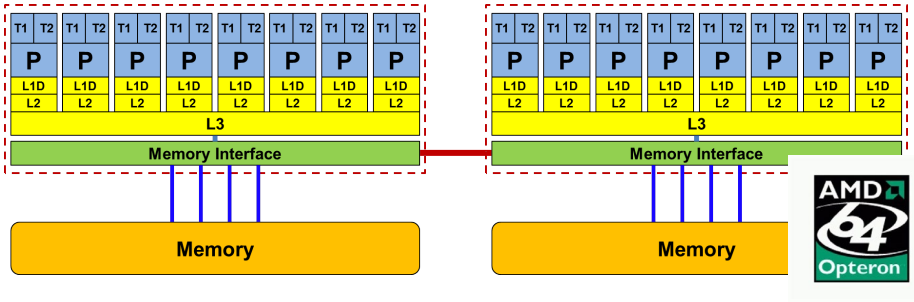
Finite-Difference 3D Navier-Stokes Solver on staggered grid

Processor	Dcavity, 128^3, 130MB
Icelake	1100.79 s
Sapphire Rapids	918.27s
M2 Pro	606.25s

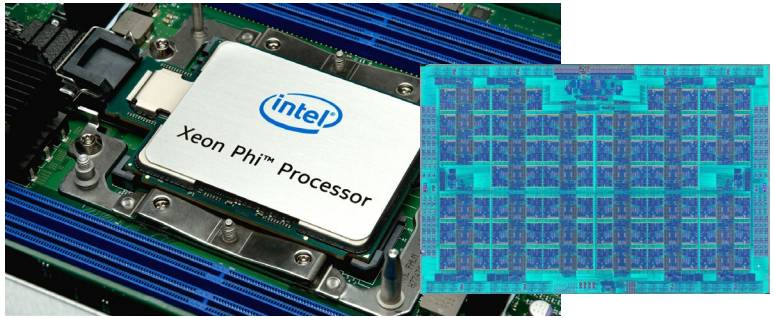
Notable success stories and failures

Itanium (2001 – 2020)

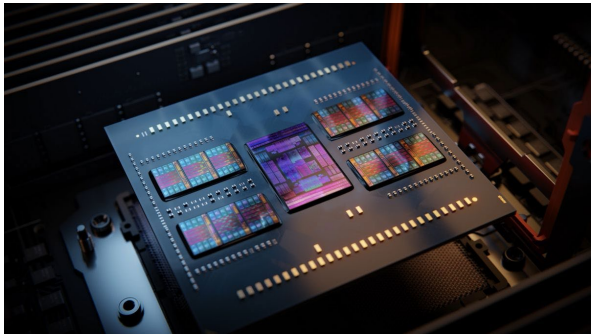
ccNUMA (2003 - now)



Xeon Phi (2015 – 2018)



Chiplets (2017 - now)



Outlook and predictions

- Extrapolation indicates that current technology is good for 2-3 more iterations

Apple's CPU development has run out of steam

Apple's improvements to the CPU cores have been marginal for years now.

- The public got used to processor improvements in terms of factors. This created a public expectation that is dissapointed if normal advances are made.
- Maybe we have to accept that performance improvements by factors was an anomaly

Outlook and predictions

- Extrapolation indicates that current technology is good for 2-3 more iteratio

Apple's CPU deve
Apple's improvemer *inal for years now.*

- The public got u n terms of factors. This created a public f normal advances are made.
- Maybe we have ovements by factors was an anomaly

Important points not covered:

- Hardware-Software Co-Design is still lacking
- Software Eco-System and social aspects
- Performance robustness and performance accessibility
- Performance degradation because of security exploits

Outlook and predictions

- Extrapolation indicates that current technology is good for 2-3 more iterations

Apple's CPU development has run out of steam

Apple's improvements to the CPU cores have been marginal for years now.

- The public got used to processor improvements in terms of factors. This created a public expectation that is dissapointed if normal advances are made.
- Maybe we have to accept that performance improvements by factors was an anomaly