How to Write Fast Numerical Code

Fall 2016

Lecture: Balance Principles, Part II

Instructor: Torsten Hoefler & Markus Püschel

TA: Salvatore Di Girolamo

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

References

- **These slides and the work is from Kenneth Czechowksi, Rich Vuduc et al., Georgia Tech**
- Kenneth Czechowski, Casey Battaglino, Chris McClanahan, Aparna Chandramowlishwaran, and Richard Vuduc. **Balance principles for algorithm-architecture co-design.** In *Proc. USENIX Wkshp. Hot Topics in Parallelism (HotPar)*, May 2011.
- Kenneth Czechowski, Chris McClanahan, Casey Battaglino, Kartik Iyer, P.-K. Yeung, Richard Vuduc. **On the communication complexity of 3D FFTs and its implications for exascale.** In *Proceedings of the ACM International Conference on Supercomputing (ICS)*, 2012.

Balance Principles II

Czechowksi et al. 2011

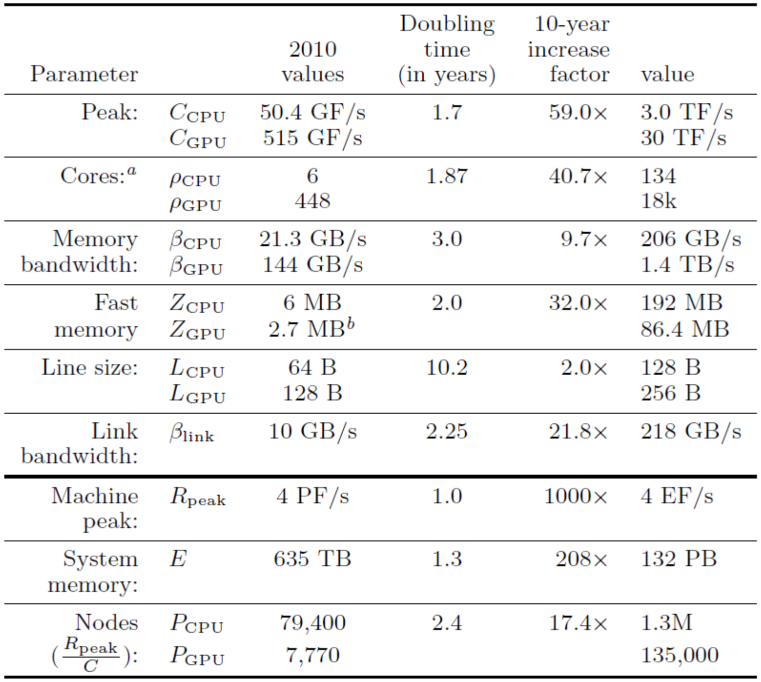
$$T_{\mathsf{mem}} \leq T_{\mathsf{comp}}$$

$$\frac{p\pi}{\beta}\left(1+\frac{\alpha\beta/\lambda}{Q/D}\right) \leq \frac{W}{Q\lambda}\left(1+\frac{p}{W/D}\right)$$

Application: Analyze Effect of HW Trends

Czechowksi et al. 2012

10 year extrapolation (2010 – 2020)



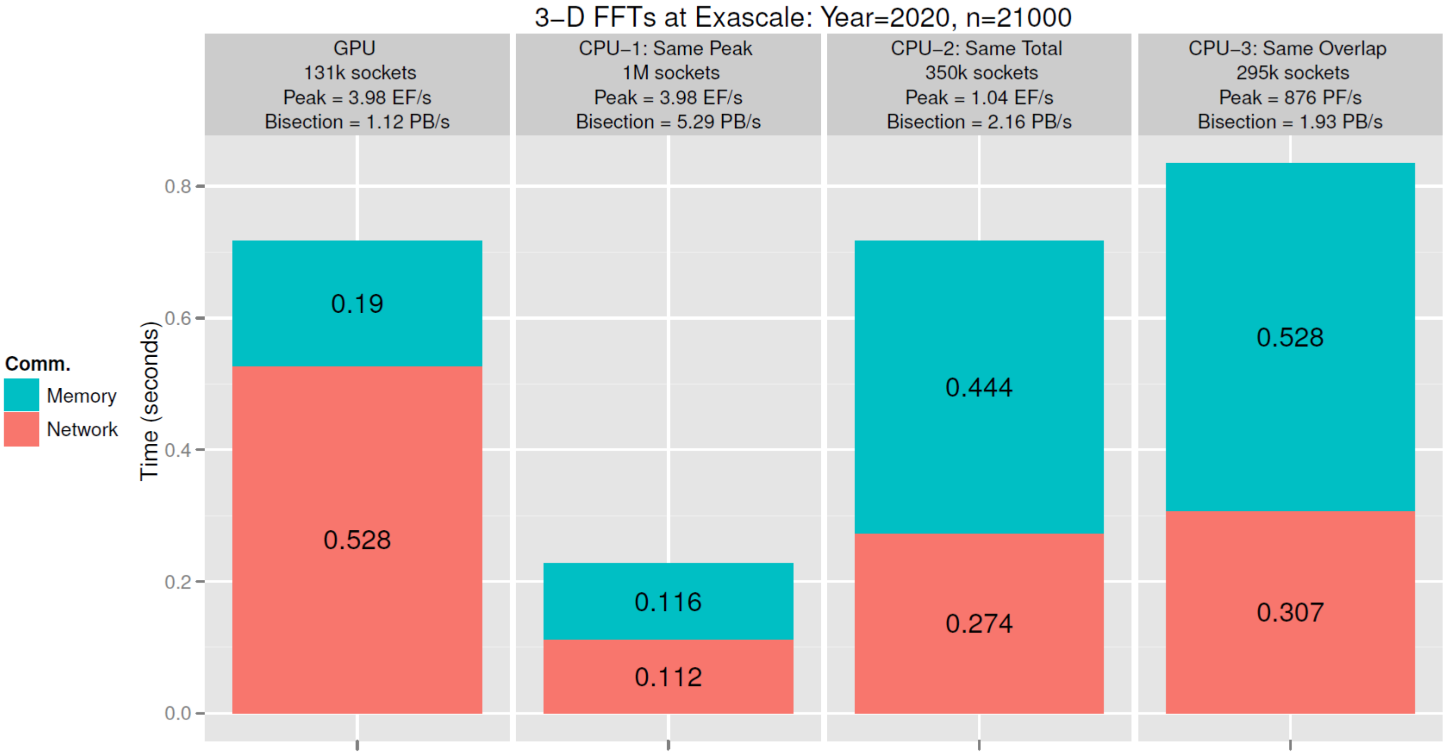
			Doubling	10-year	
		2010	time	increase	
Parameter		values	(in years)	factor	value
Peak:	C_{CPU} C_{GPU}	50.4 GF/s 515 GF/s	1.7	59.0×	3.0 TF/s 30 TF/s
Cores:[a]	ρ_{CPU} ρ_{GPU}	6 448	1.87	40.7×	134 18k
Memory bandwidth:	β_{CPU} β_{GPU}	21.3 GB/s 144 GB/s	3.0	9.7×	206 GB/s 1.4 TB/s
Fast memory	Z_{CPU} Z_{GPU}	6 MB 2.7 MB[b]	2.0	32.0×	192 MB 86.4 MB
Line size:	L_{CPU} L_{GPU}	64 B 128 B	10.2	2.0×	128 B 256 B
Link bandwidth:	β_{link}	10 GB/s	2.25	21.8×	218 GB/s
Machine peak:	R_{peak}	4 PF/s	1.0	1000×	4 EF/s
System memory:	E	635 TB	1.3	208×	132 PB
Nodes $(\frac{R_{\text{peak}}}{C})$:	P_{CPU} P_{GPU}	79,400 7,770	2.4	17.4×	1.3M 135,000

Application: Analyze Effect of HW Trends

Czechowksi et al. 2012

3D-FFT in 2020:

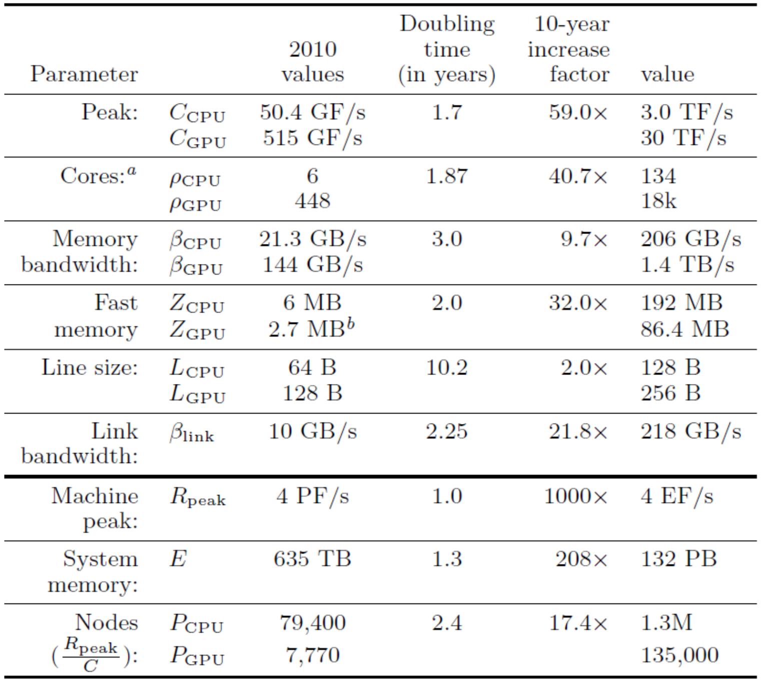
Faster on CPU or GPU?



Application: Analyze Effect of HW Trends

Czechowksi et al. 2012

10 year extrapolation (2010 – 2020)

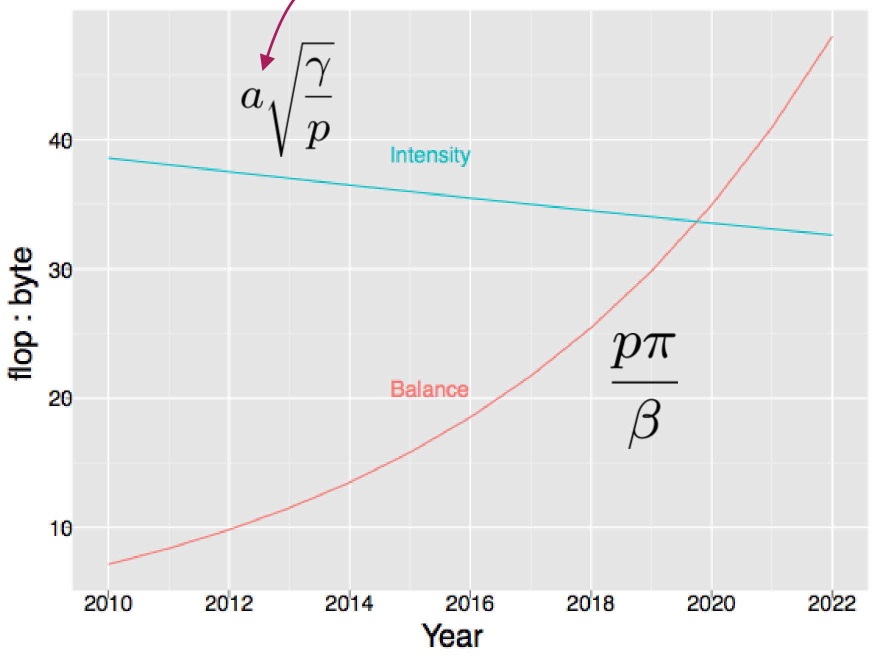


Parameter		2010 values	Doubling time (in years)	10-year increase factor	value
Peak:	C_{CPU} C_{GPU}	50.4 GF/s 515 GF/s	1.7	59.0×	3.0 TF/s 30 TF/s
Cores:[a]	ρ_{CPU} ρ_{GPU}	6 448	1.87	40.7×	134 18k
Memory bandwidth:	β_{CPU} β_{GPU}	21.3 GB/s 144 GB/s	3.0	9.7×	206 GB/s 1.4 TB/s
Fast memory	Z_{CPU} Z_{GPU}	6 MB 2.7 MB[b]	2.0	32.0×	192 MB 86.4 MB
Line size:	L_{CPU} L_{GPU}	64 B 128 B	10.2	2.0×	128 B 256 B
Link bandwidth:	β_{link}	10 GB/s	2.25	21.8×	218 GB/s
Machine peak:	R_{peak}	4 PF/s	1.0	1000×	4 EF/s
System memory:	E	635 TB	1.3	208×	132 PB
Nodes $\left(\frac{R_{\text{peak}}}{C}\right)$:	P_{CPU} P_{GPU}	79,400 7,770	2.4	17.4×	1.3M 135,000

Matrix-multiplication on GPU

$$\frac{p\pi}{\beta} \le O\left(\sqrt{\frac{\gamma}{p}}\right)$$

through measurements



Even Matmult on GPU could become memory bound!

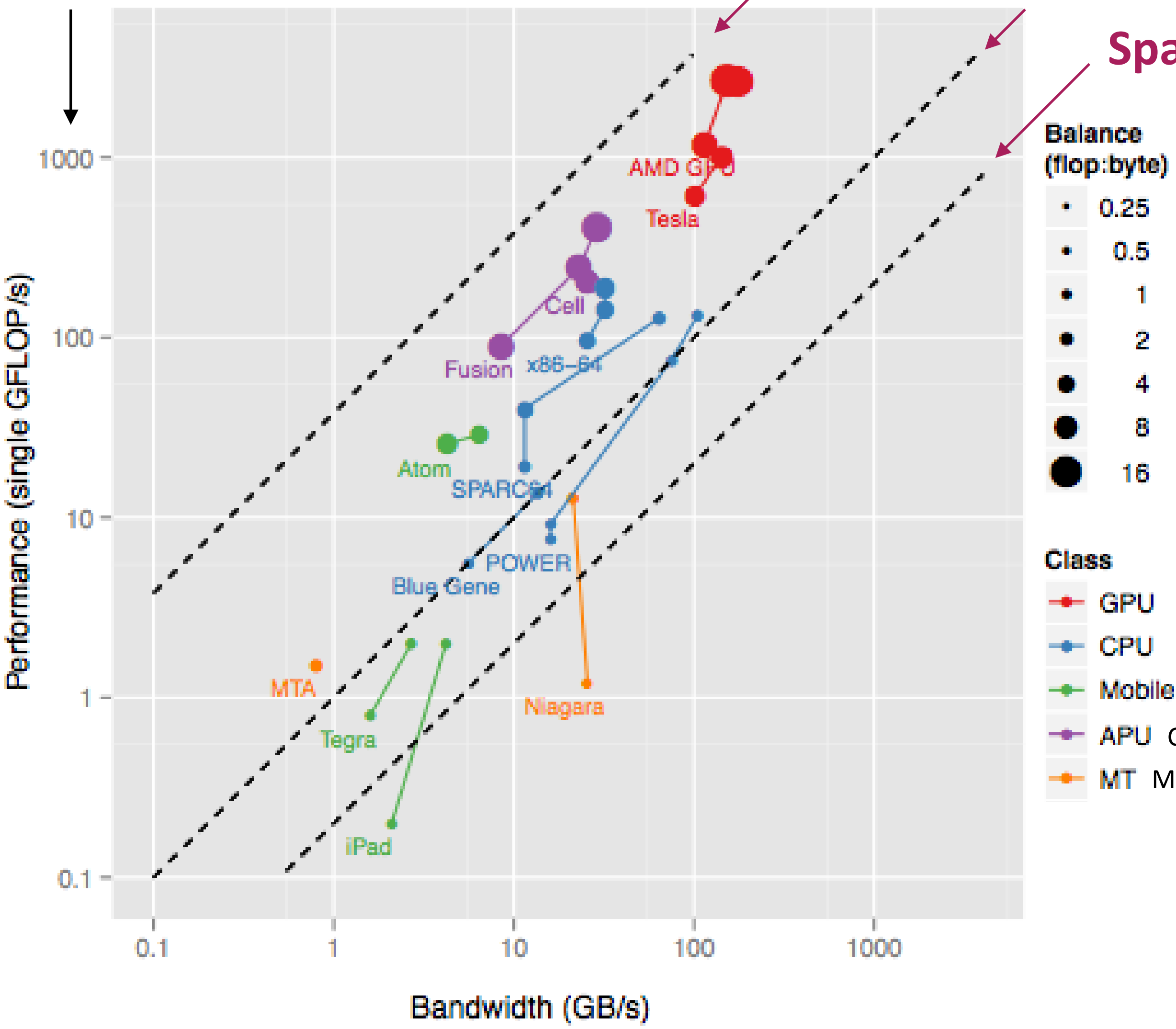
Approximate operational intensity

Matmult

FFT

Sparse Matvec

$p\pi$



CPU+GPU

Multithreaded (old)

β

1. More powerful: less balance
2. Build large scale with low power processors