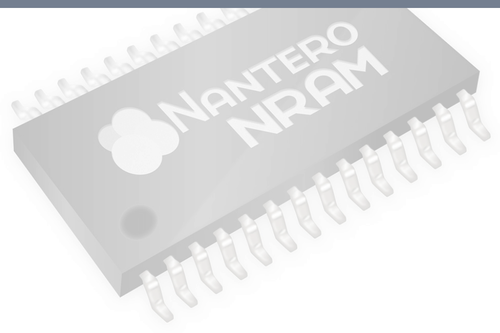
Neural networks,

Disclaimer:

The contents of this talk do **not** necessarily reflect the quality, rigor, and integrity of my lab's actual technical work.

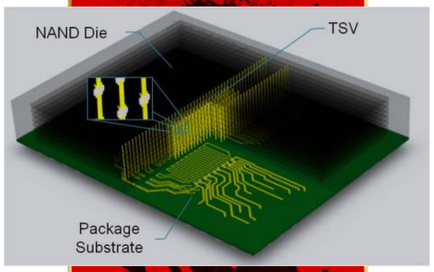


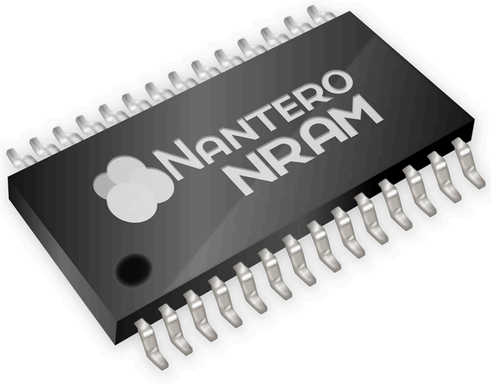
Richard (Rich) Vuduc

Georgia Tech College of Computing
Computational Science and Engineering

hpcgarage

hpcgarage.org/salishan





Neural networks, human brains, and "The Memoristas"

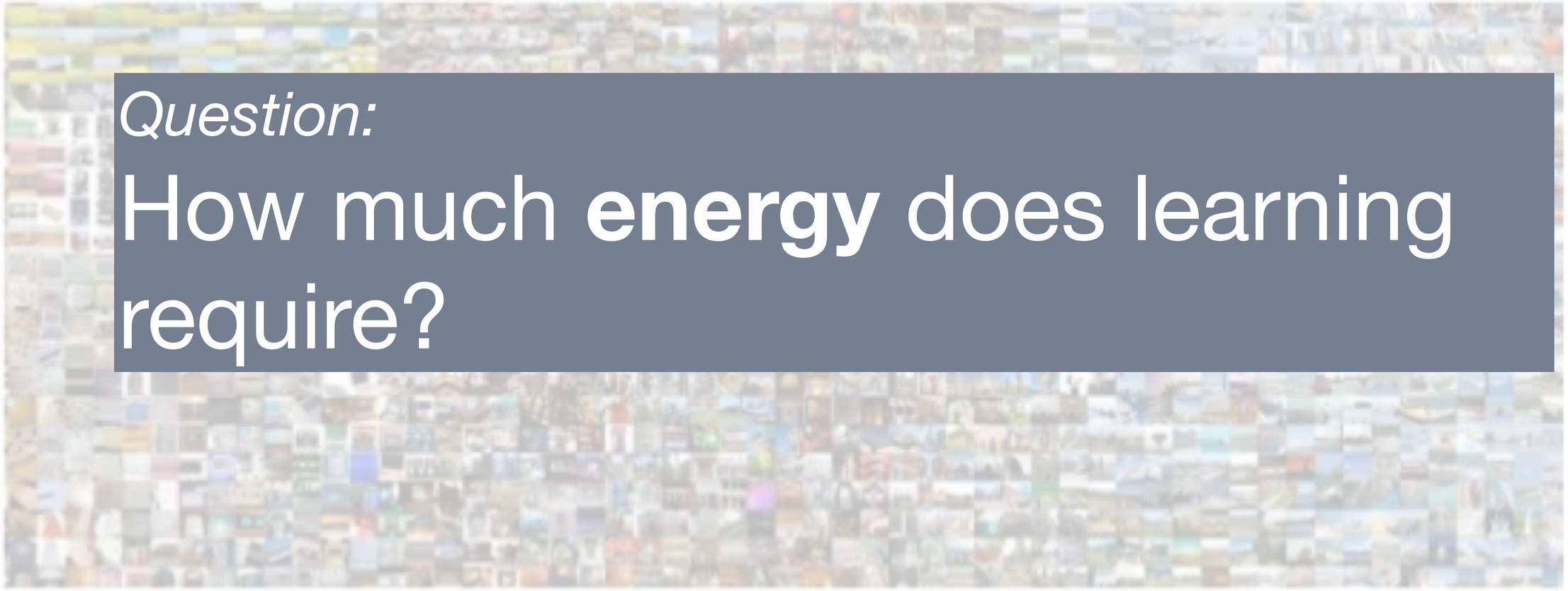
Richard (Rich) Vuduc

Georgia Tech College of Computing
Computational Science and Engineering

hpcgarage

ImageNet Dataset

IMAGENET



Question:

How much **energy** does learning require?

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*. [web]

ImageNet Dataset

IMAGENET

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*. [web]

(4 GPUs)
x (250 Watts / GPU)
x (1 week)
~ 0.6 billion Joules

(4 GPUs)
x (250 Watts / GPU)
x (1 week)
~ 0.6 billion Joules

(1 brain)
x (20 Watts / brain)
x (1 year)
~ 0.6 billion Joules

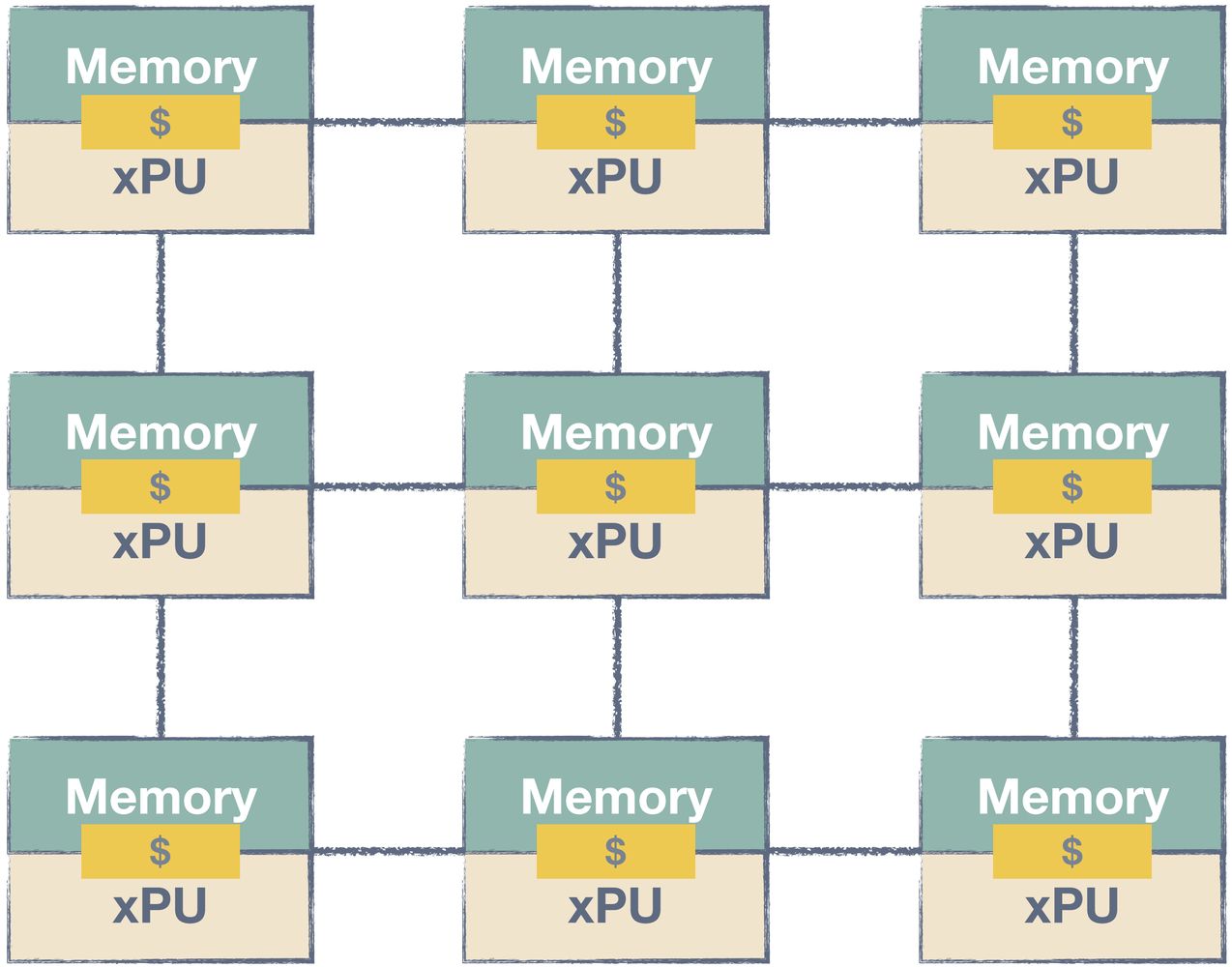
(4 GPUs)
x (250 Watts / GPU)
x (1 week)
~ 0.6 billion Joules

??

(1 brain)
x (20 Watts / brain)
x (1 year)
~ 0.6 billion Joules

The Iron Law

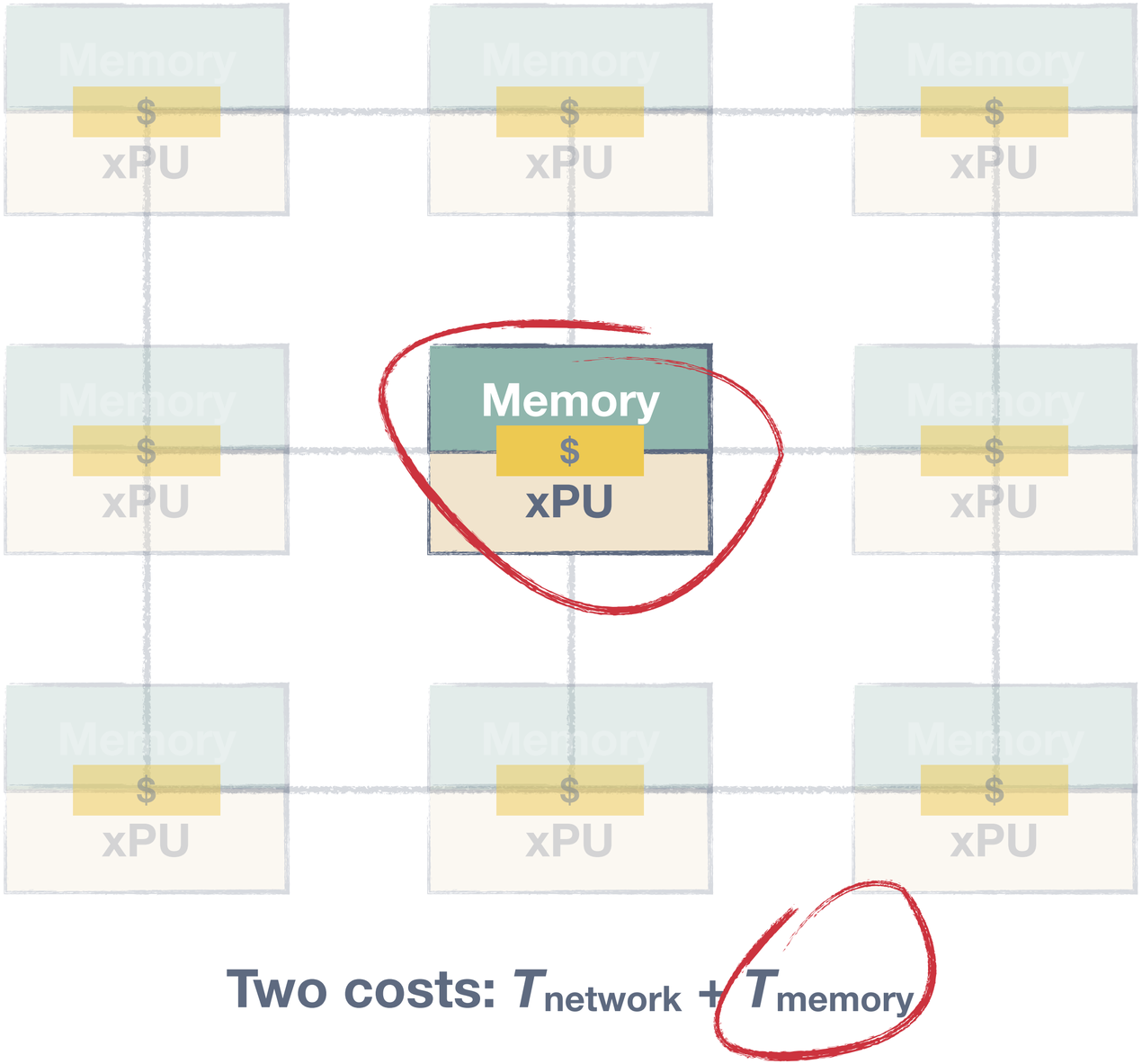
of parallel and distributed communication costs



Two costs: $T_{network} + T_{memory}$

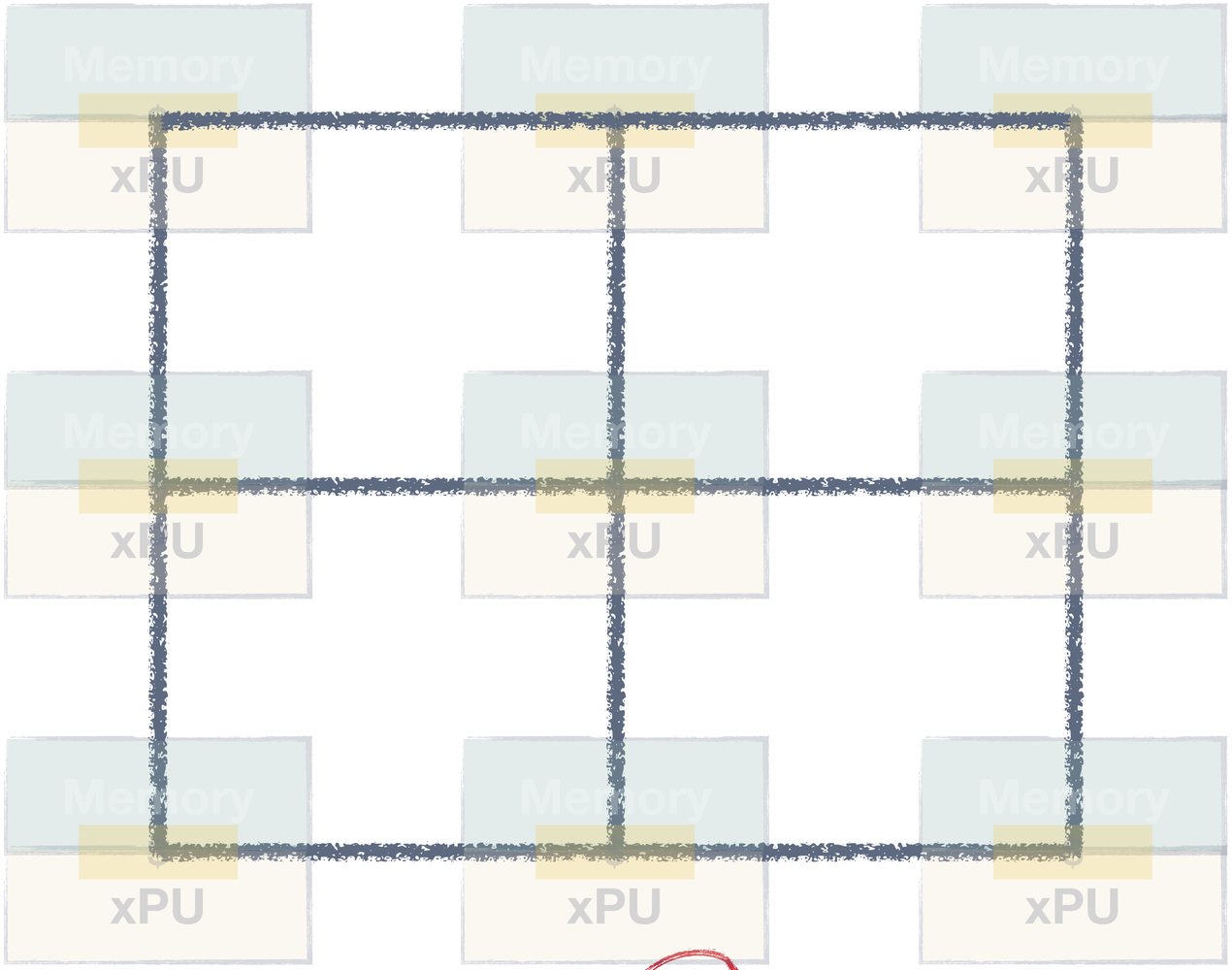
The Iron Law

of parallel and distributed communication costs



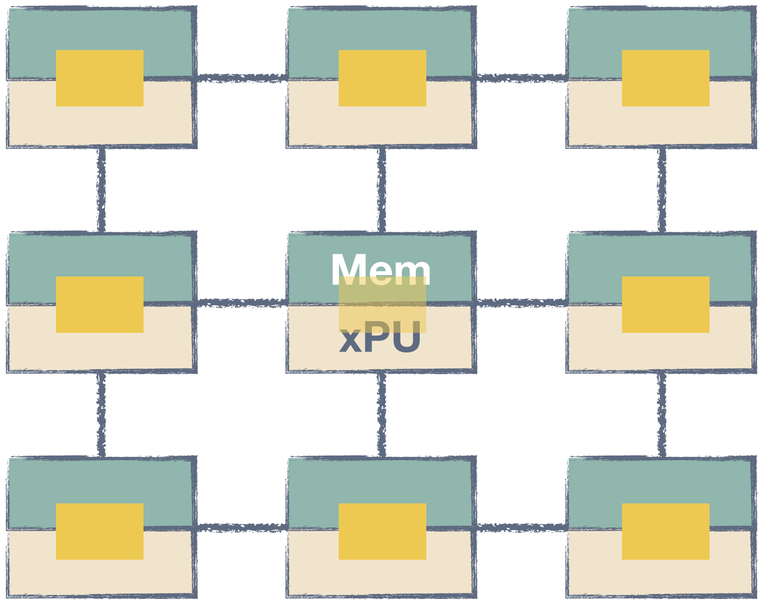
The Iron Law

of parallel and distributed communication costs



Two costs: $T_{network}$ + T_{memory}

Two costs: $T_{\text{network}} + T_{\text{memory}}$



A fixed comm-bound algorithm + fixed problem size + fixed machine peak:

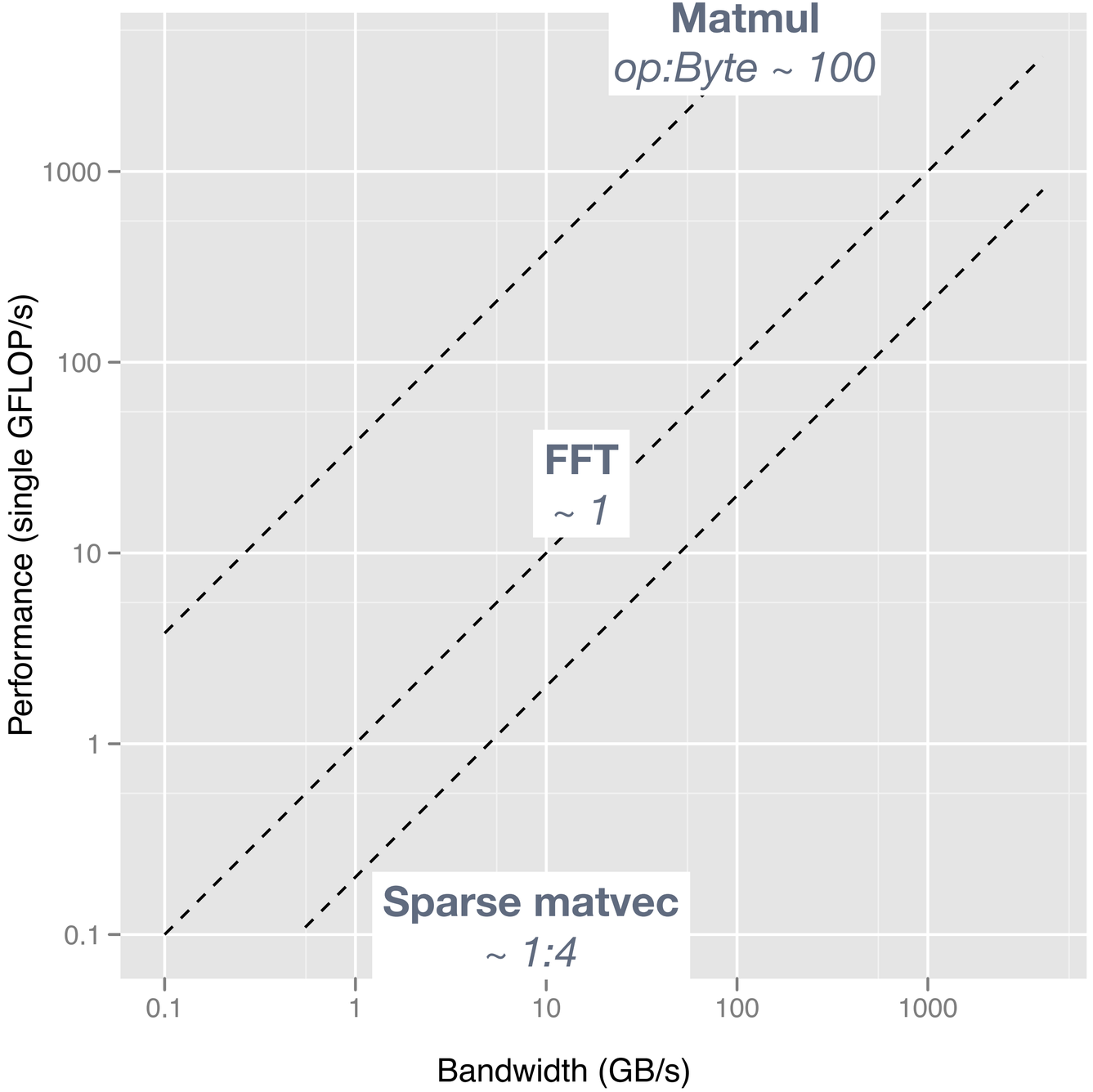
$$\text{Let } B \equiv \frac{\text{xPU flop/s}}{\text{xPU memory bandwidth}}$$

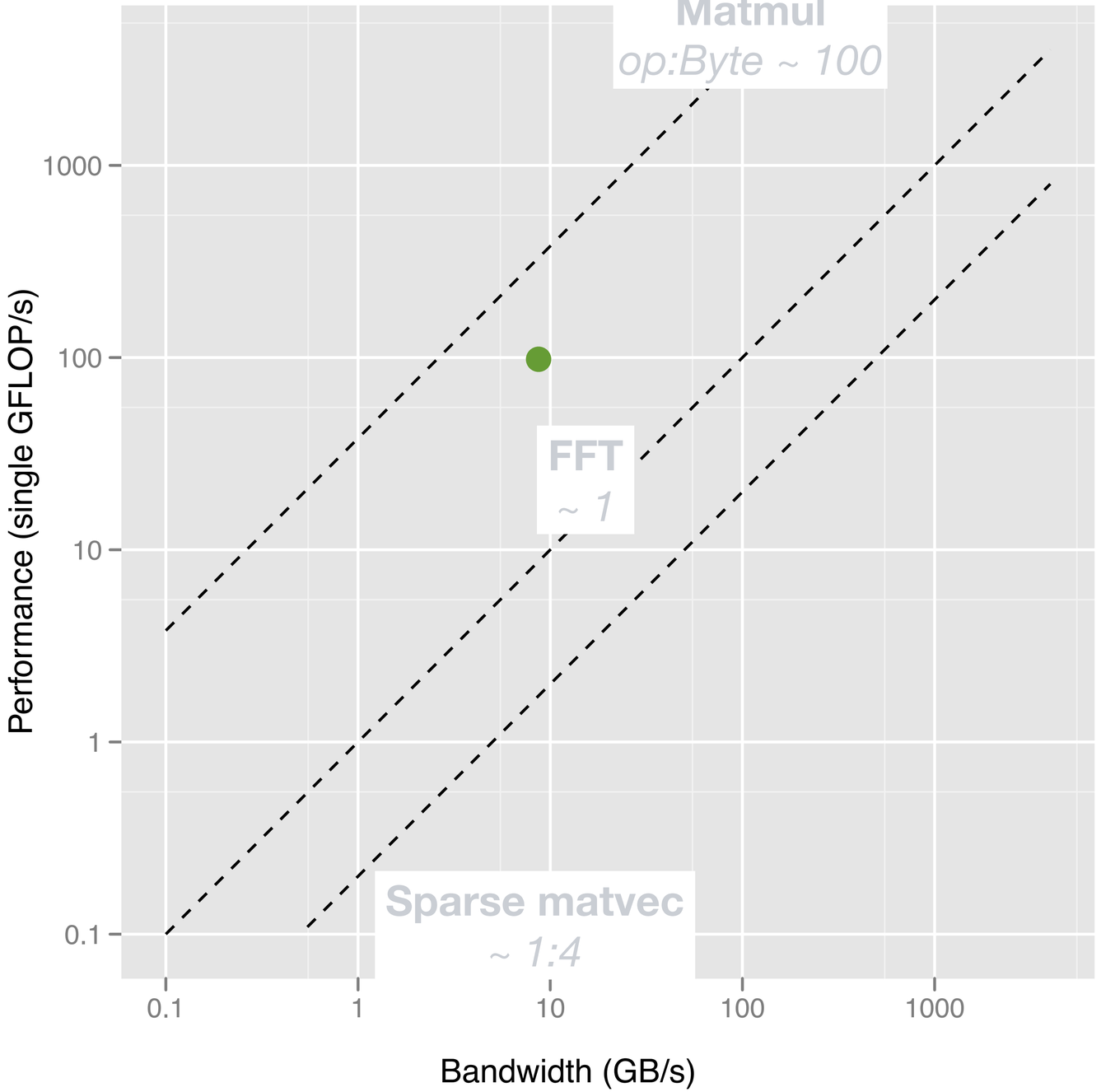
$$T_{\text{mem}} \sim \mathcal{O}(B)$$

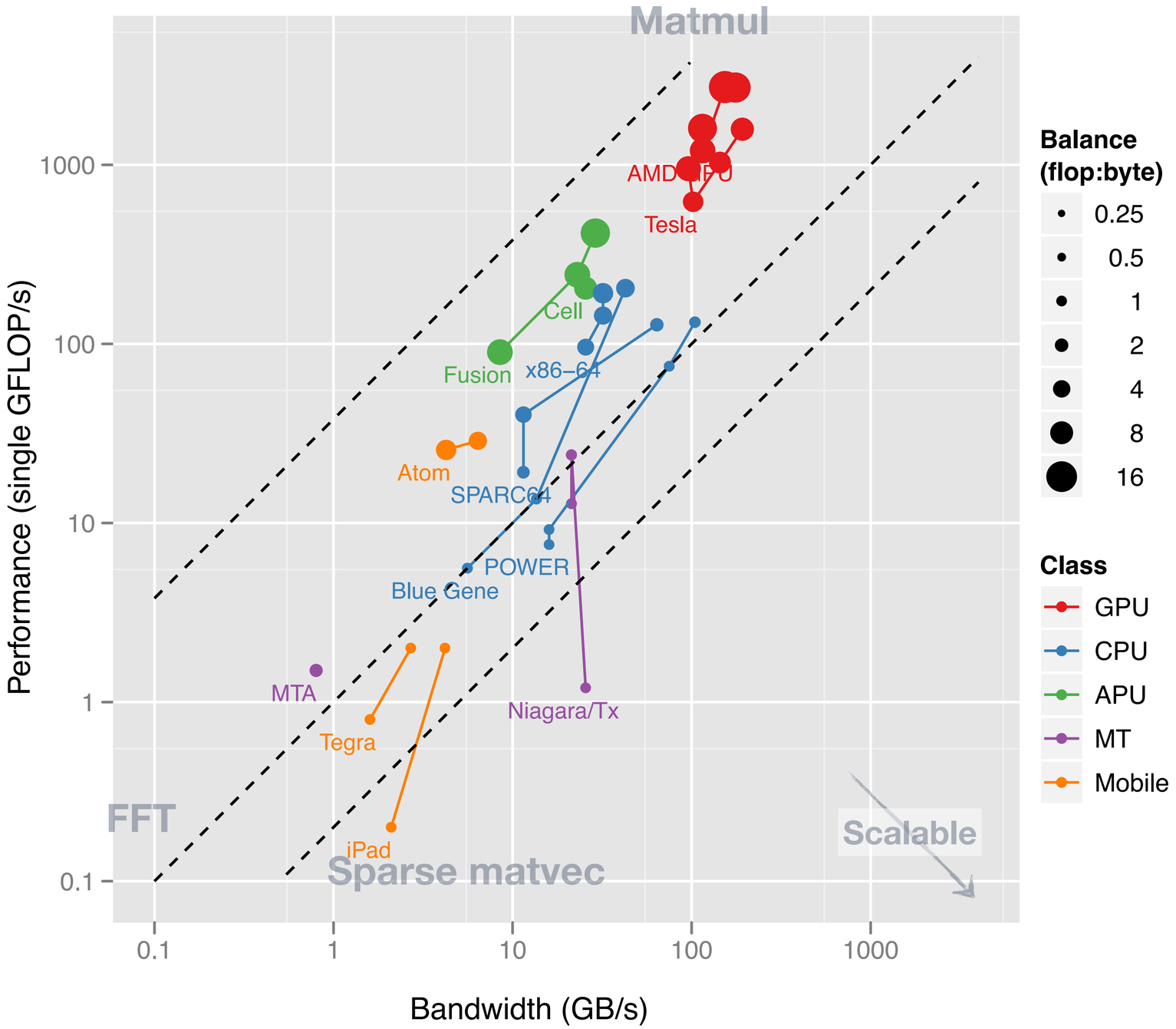
$$T_{\text{net}} \sim \mathcal{O}\left(B^{\kappa} \cdot \frac{\beta_{\text{mem}}^{\kappa}}{\beta_{\text{link}}}\right), \qquad \kappa = (d-1)/d$$

(d-dimensional torus)

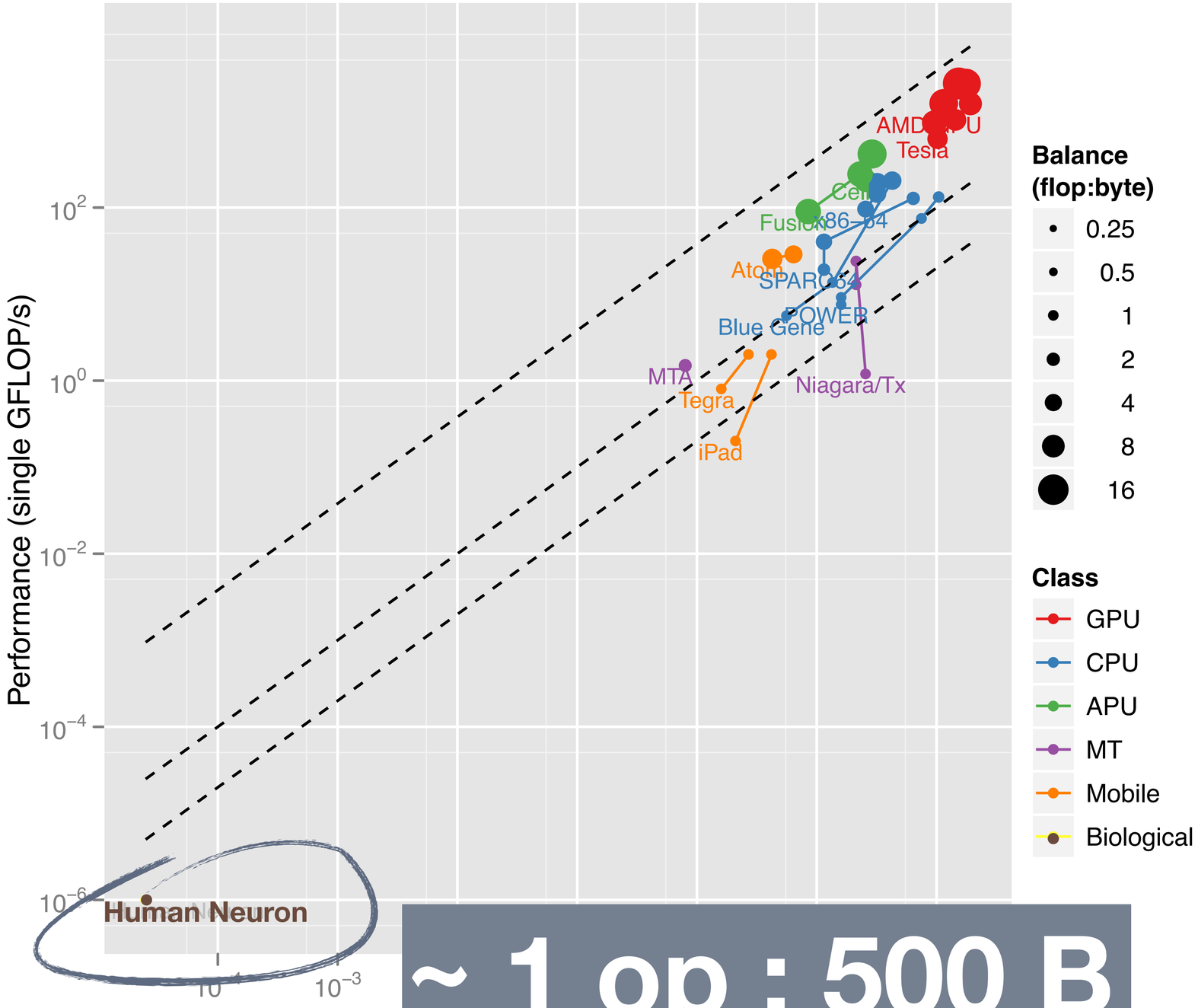
B = "***balance***" — Kung (1986); Hillis (1987); Callahan & Kennedy (1988); Blelloch, Maggs, Miller (1994); McCalpin (1995); Williams et al. (2009)
See also: Czechowski et al. (ICS'2012); Young & Vuduc (PMES'2016)


Matmul
op:Byte ~ 100
FFT
~ 1
Sparse matvec
~ 1:4
Performance (single GFLOP/s)
1000
100
10
1
0.1
0.1
1
10
100
1000
Bandwidth (GB/s)


Matmul
op:Byte ~ 100
FFT
~ 1
Sparse matvec
~ 1:4
Performance (single GFLOP/s)
1000
100
10
1
0.1
0.1
1
10
100
1000
Bandwidth (GB/s)

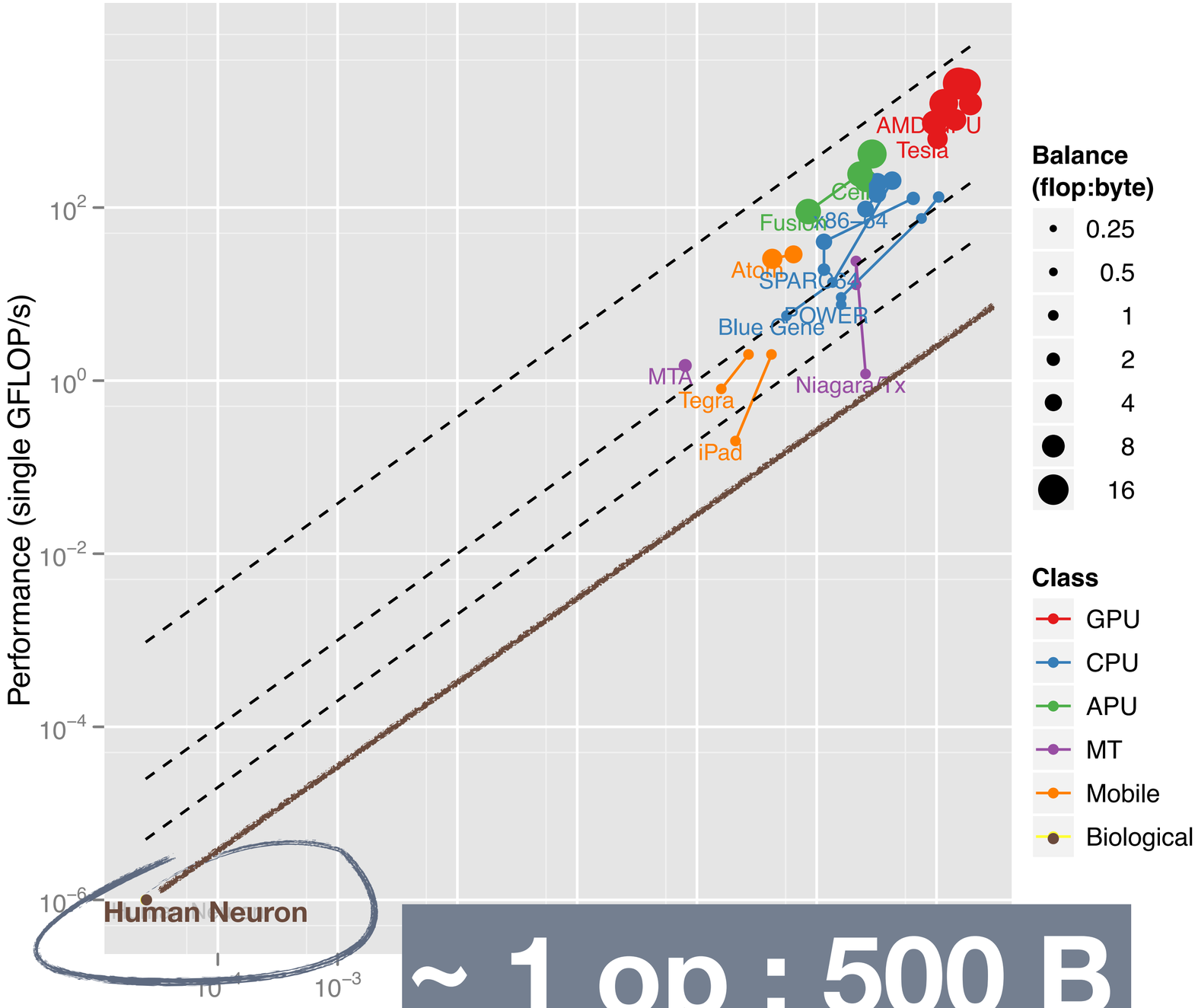


Data circa 2012



~ 1 op : 500 B

http://www.jetpress.org/volume1/moravec.pdf



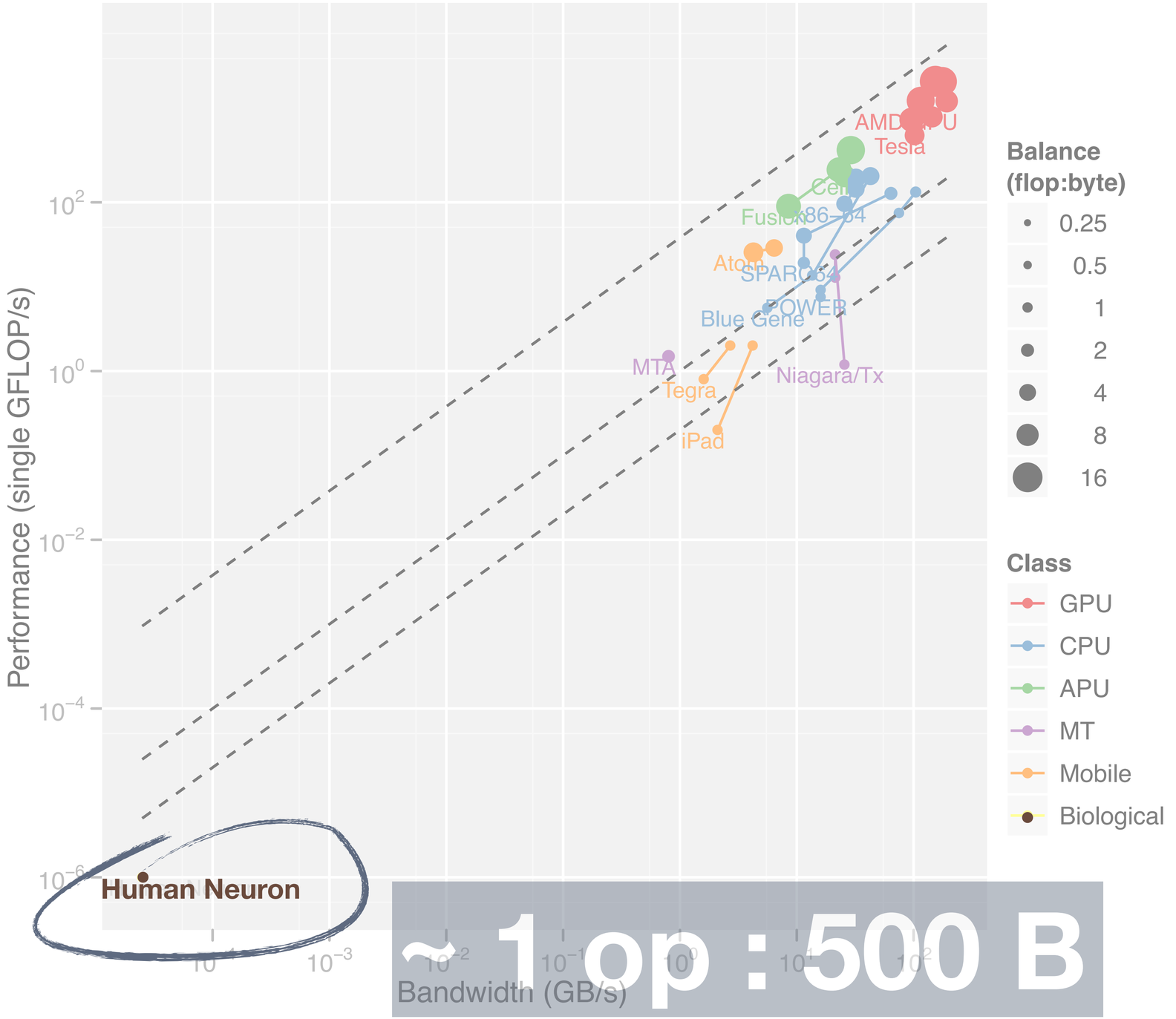
~ 1 op : 500 B

http://www.jetpress.org/volume1/moravec.pdf

Communication capability dominates that of computation — huh?

This ratio almost *appears* nonsensical as we think of computation **today**. Is it?

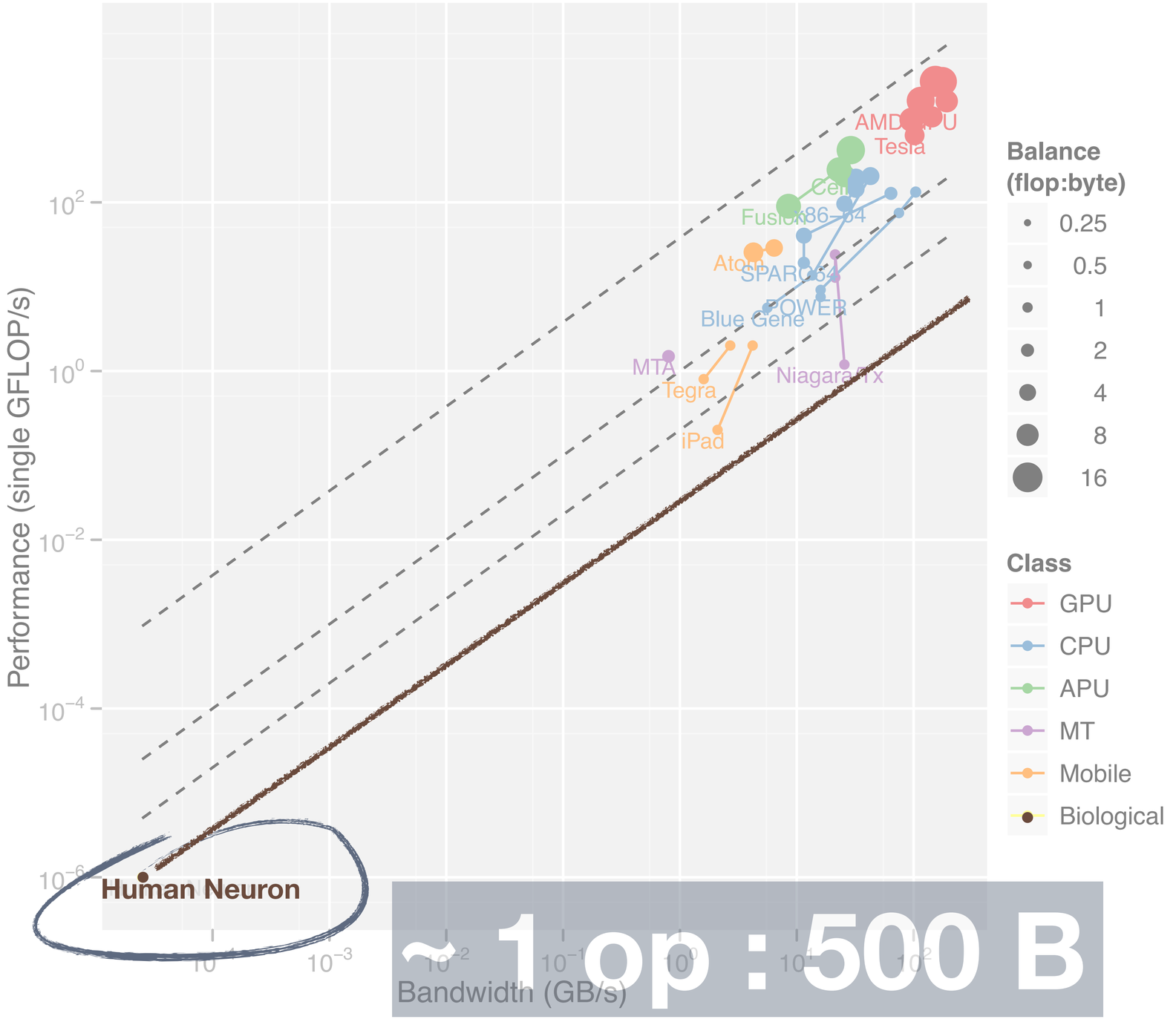
Interpretations: New computational models are needed, perhaps that **reduce**, **sample**, and **approximate** judiciously, or where the mere act of moving bits **is** the computation.



Communication capability dominates that of computation — huh?

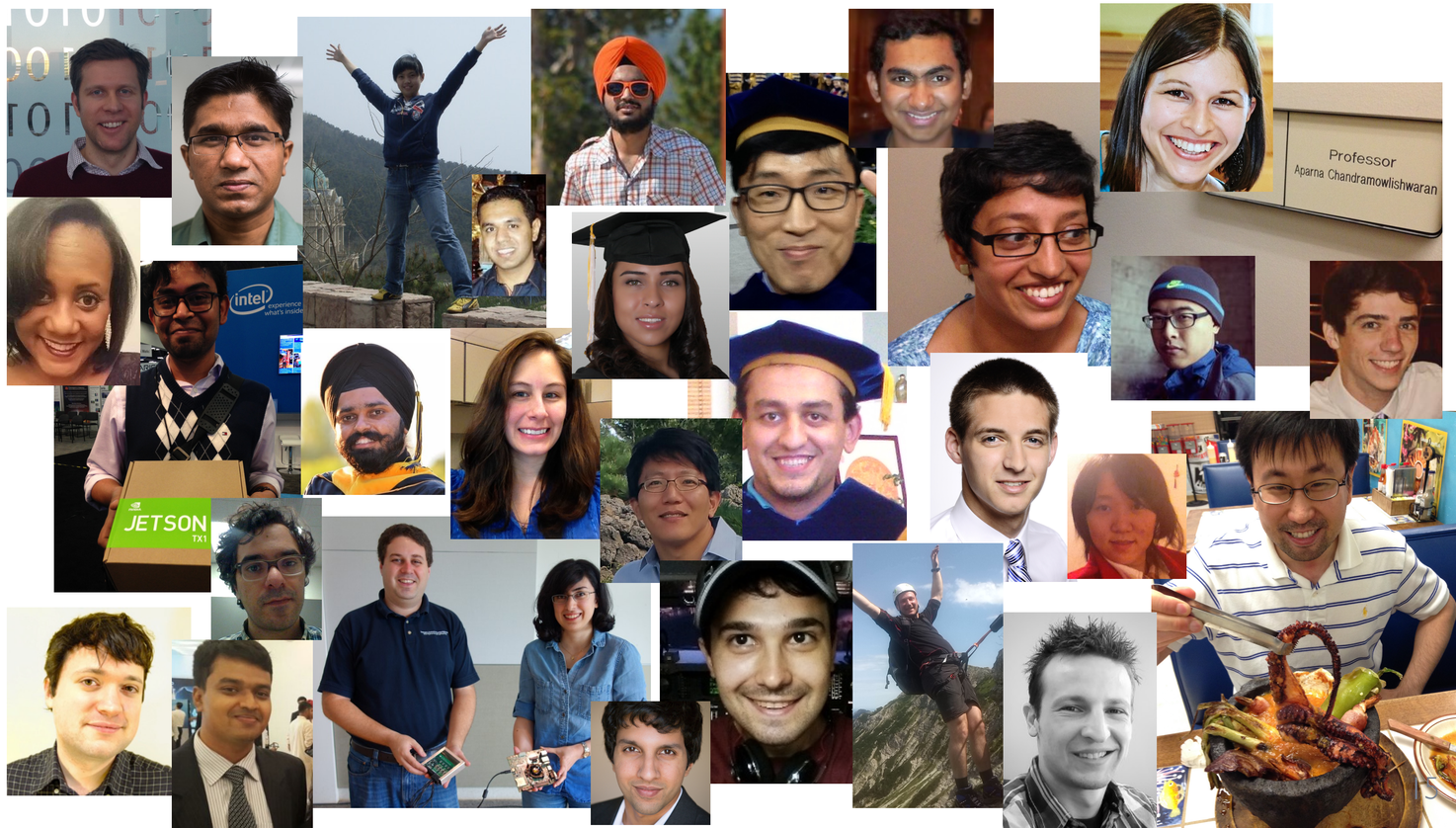
This ratio almost *appears* nonsensical as we think of computation **today**. Is it?

Interpretations: New computational models are needed, perhaps that **reduce**, **sample**, and **approximate** judiciously, or where the mere act of moving bits **is** the computation.



hpcgarage.org/salishan

hpcgarage



Rich Vuduc

High-performance computing,
scalable parallel algorithms & software

Georgia Tech College of Computing
Computational Science and Engineering