Approximate Computing and Microfluidic Cooling for Enhanced Machine Learning

Hardik Sharma

†William Wahby

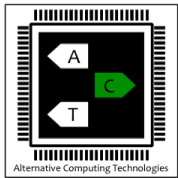
†Thomas Sarvey

†Muhannad S. Bakir

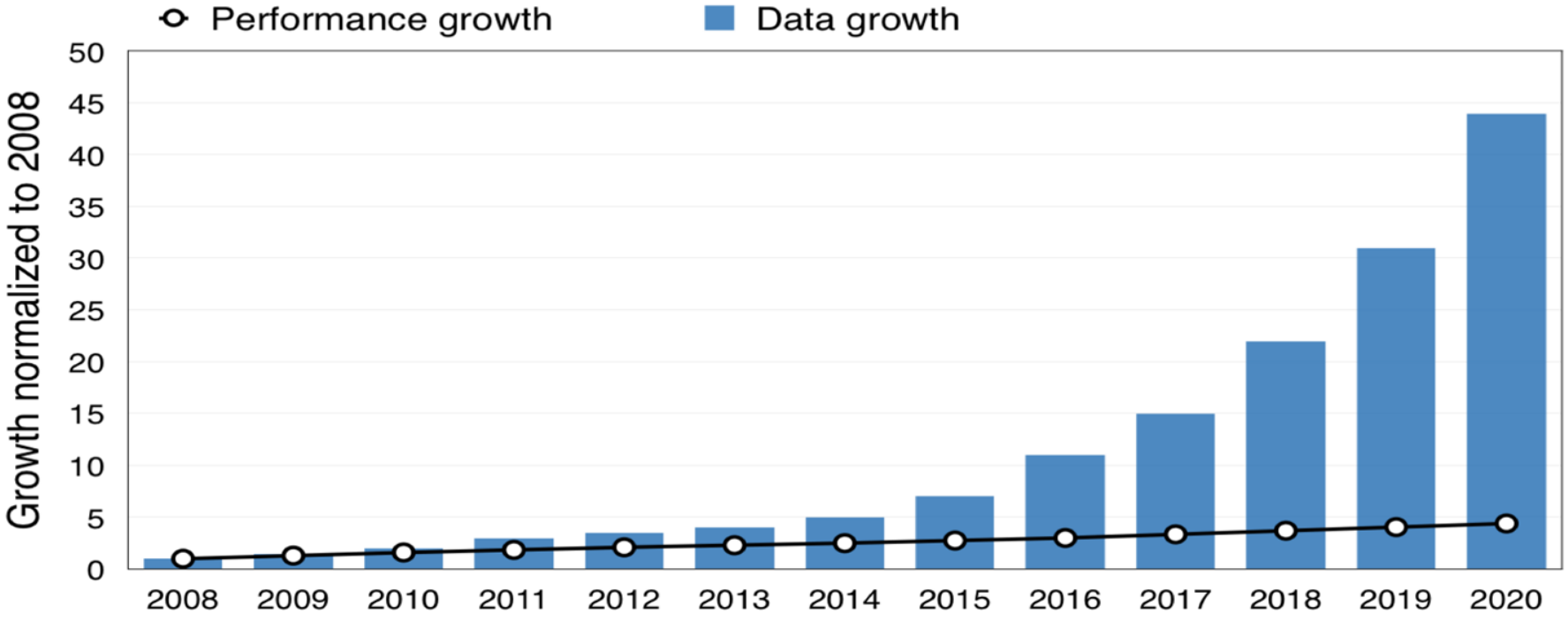
Hadi Esmaeilzadeh

Alternative Computing Technologies (**ACT**) Lab
Georgia Institute of Technology

†Integrated 3D Systems (**I3DS**) Group
Georgia Institute of Technology

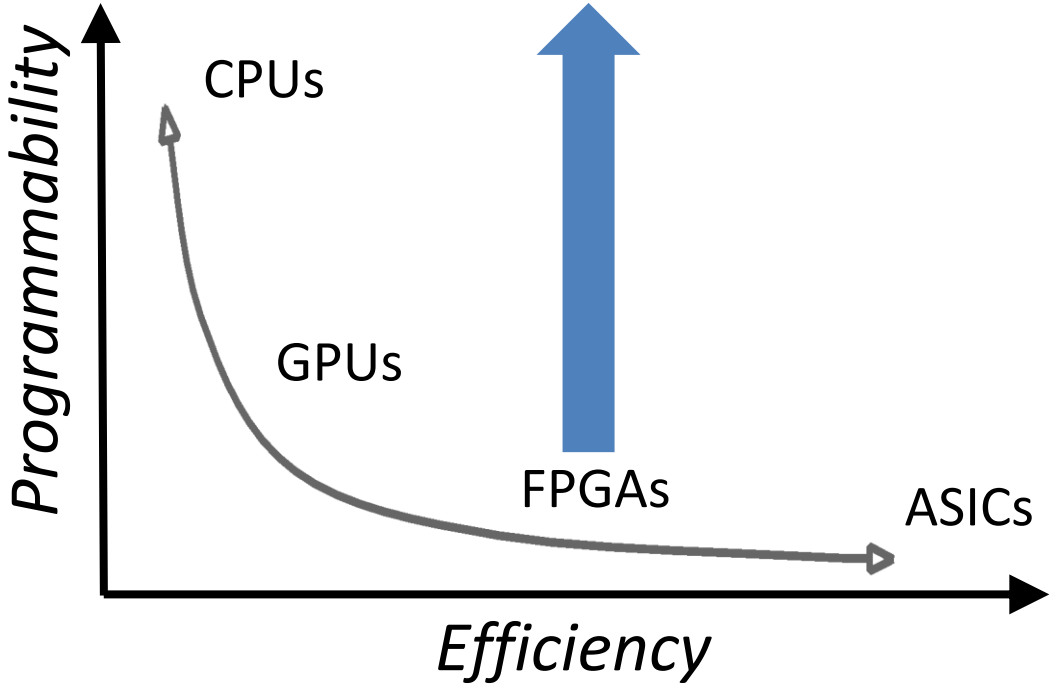


Data growth vs. Performance

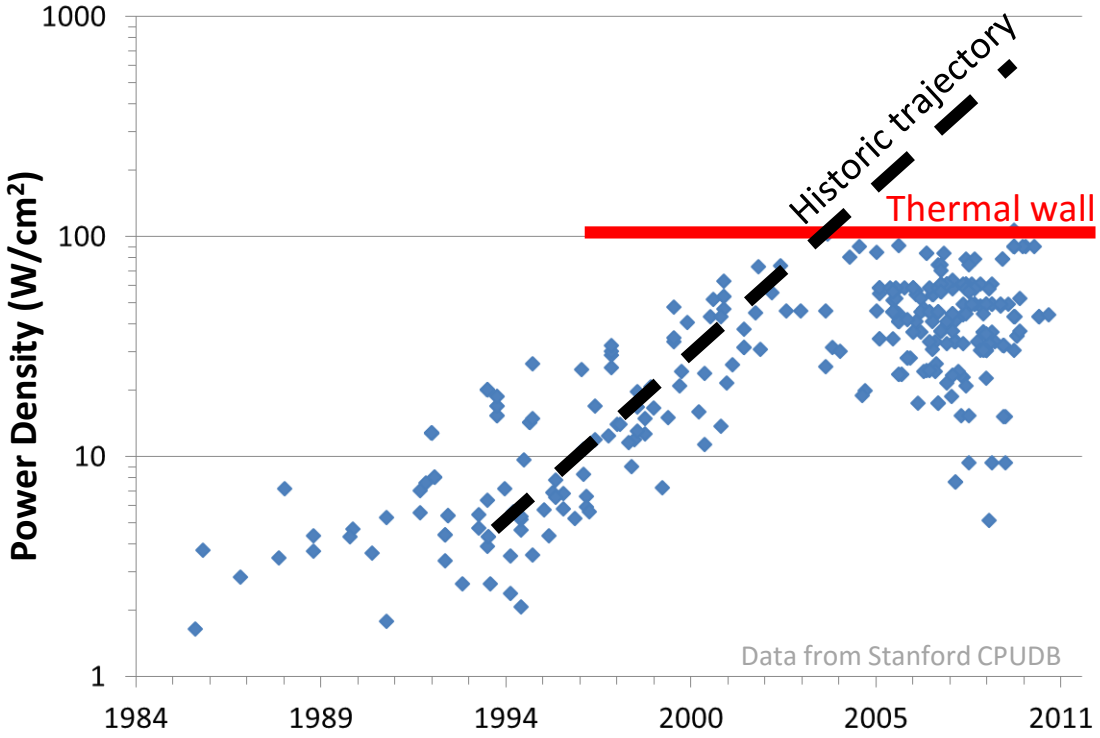


Data growth trends: IDC's Digital Universe Study, December 2012
Performance growth trends: Esmaeilzadeh et al, “Dark Silicon and the End of Multicore Scaling,” ISCA 2011

Programmability vs. Efficiency

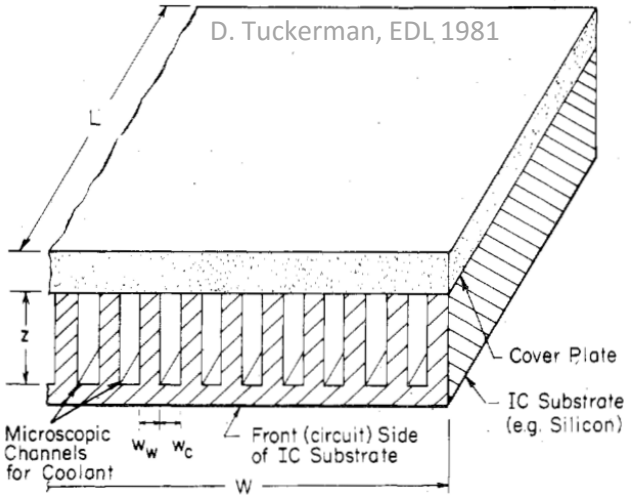


Heat limits system performance

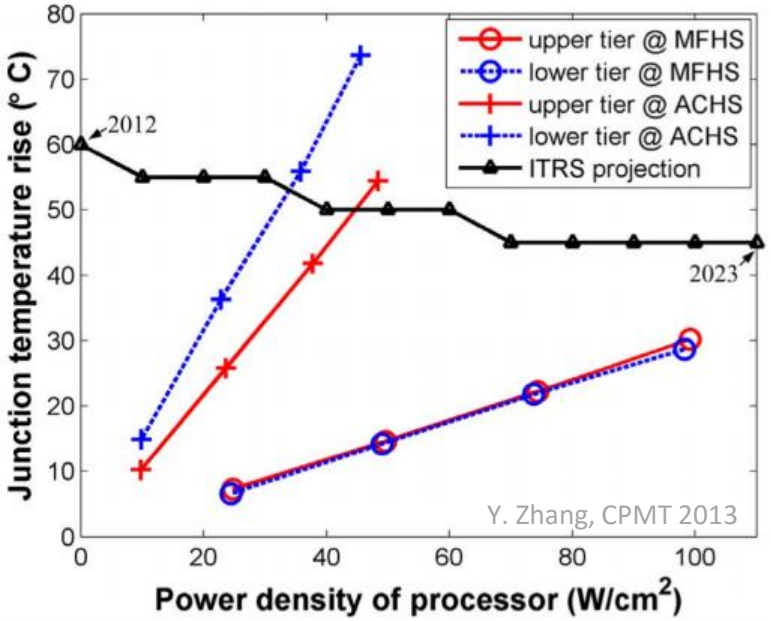


Better cooling improves performance

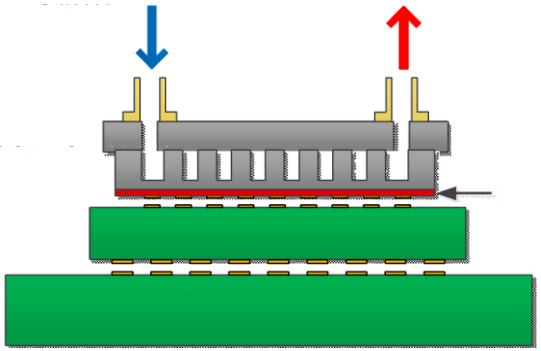
First Proposed in 1981



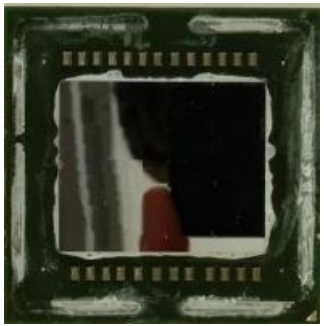
Non-functional characterization in 2013



Microfluidically-cooled Stratix V FPGA

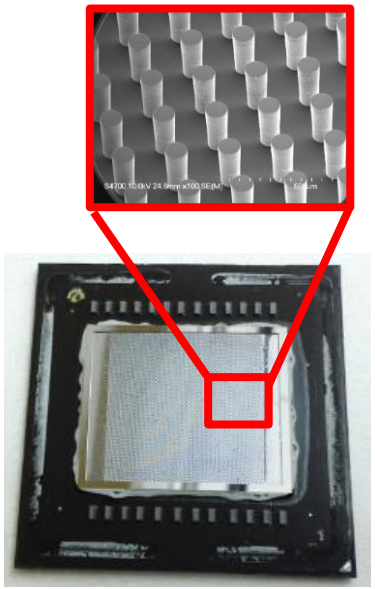


Original Die



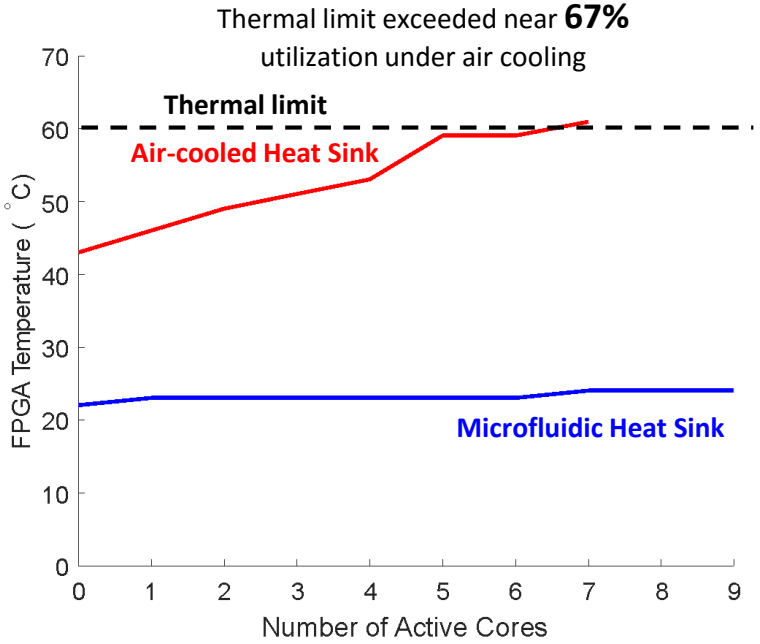
Delidded die

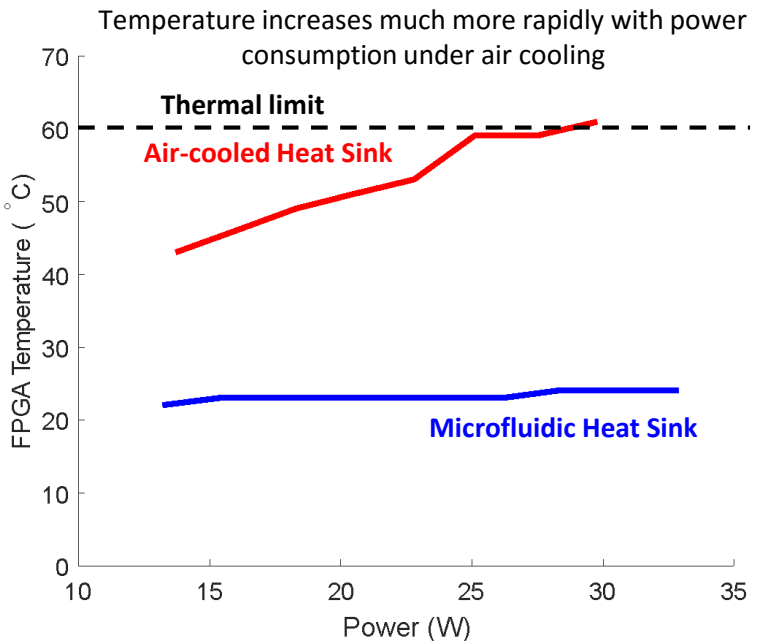
Micropinfin closeup



Etched Microfluidic Heat Sink

Microfluidically-cooled Stratix V FPGA

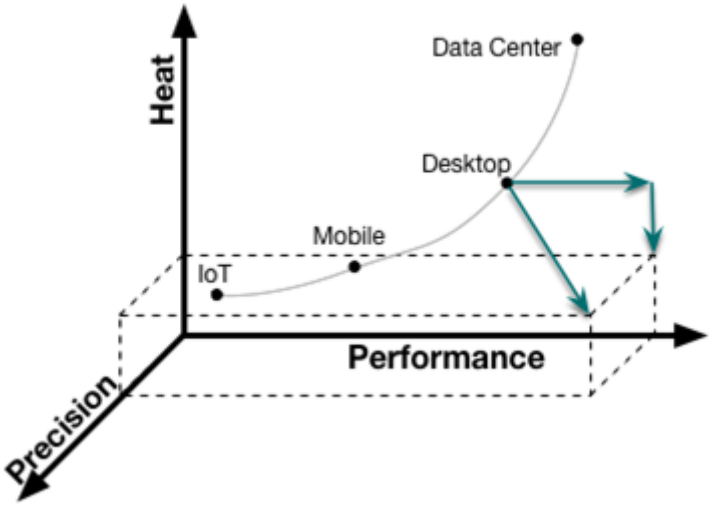




Junction-to-ambient $R_{th} \approx 0.08°C/W$

Nominally expect **only 40 °C increase** over ambient at **500W** power dissipation

Approximate Computing for FPGA Acceleration

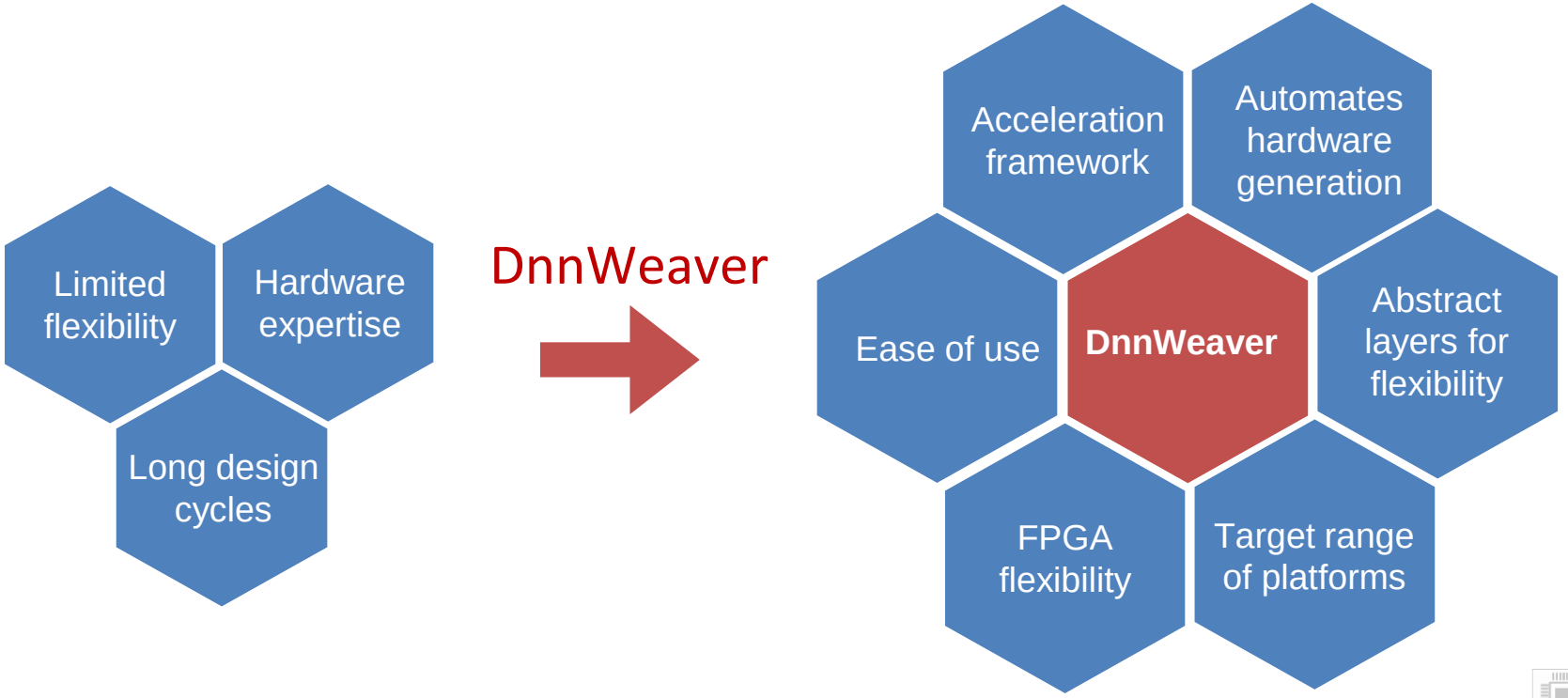


Deep Neural Networks have high tolerance to approximation.
(DeepCompression ICLR2016)

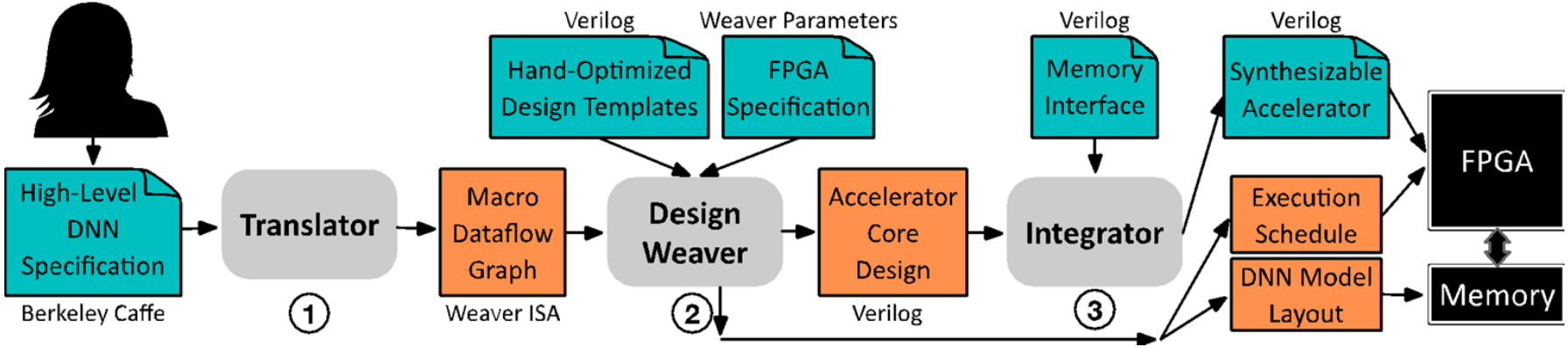
Relaxing Precision will yield higher performance.

- Increased parallelism through reduced resource usage.
- Reduce bandwidth.

Challenges in Hardware Acceleration



Compilation flow



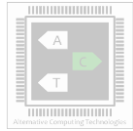
We present a comprehensive framework for accelerating DNNs from high-level abstractions

Benchmarks

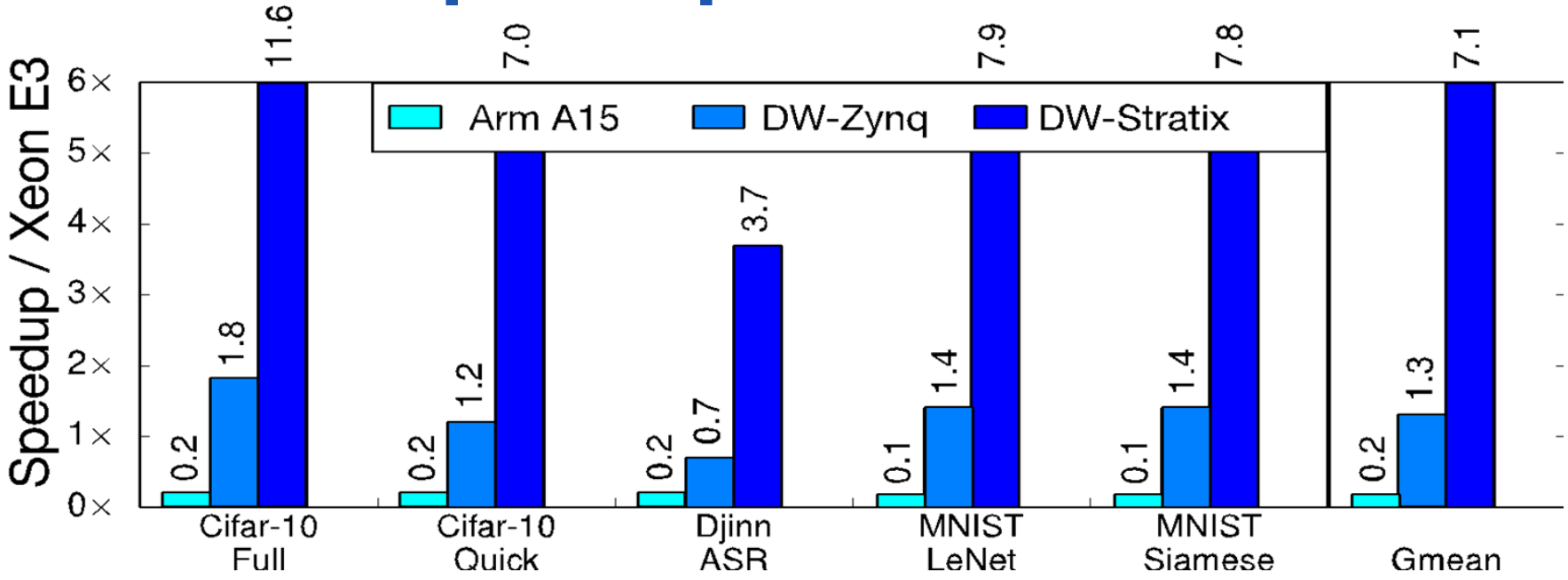
LENET	Character recognition	C->P->C->P->I->A->I
Siamese	Character recognition	C->P->C->P->I->A->I->A
CIFAR 10 -Quick	Object Recognition	C->P->A->N->C->A->P->N->C->A->P->I
CIFAR 10 -Full	Object Recognition	C->P->A->C->A->P->C->A->P->I->I
DJINN ASR	Speech to text Decoder	I->A->I->A->I->A->I->A->I->A->I->A->I

Evaluated Platforms

FPGA	Altera Stratix V TDP:25W $6999	Xilinx Zynq 7000 ZC702 TDP: 2W $129	
CPU	Intel Xeon E3-1276 V3 TDP: 84W $339	ARM Cortex 15 TDP: 5W $191	
GPU	Tegra K1 GPU TDP: 10 W $191	GeForce GTX 650 Ti TDP: 110 $150	Tesla K40 TDP: 235 W $5499

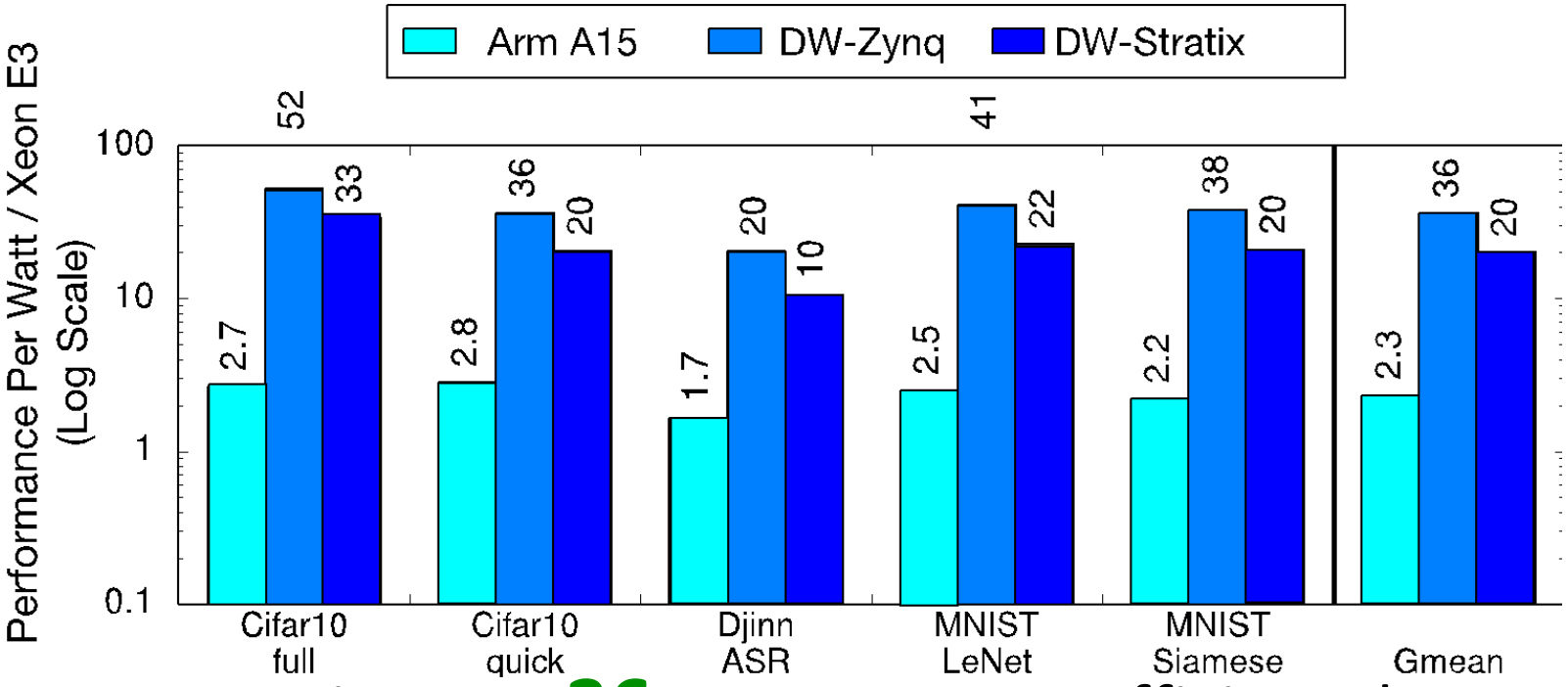


Speedup over CPUs



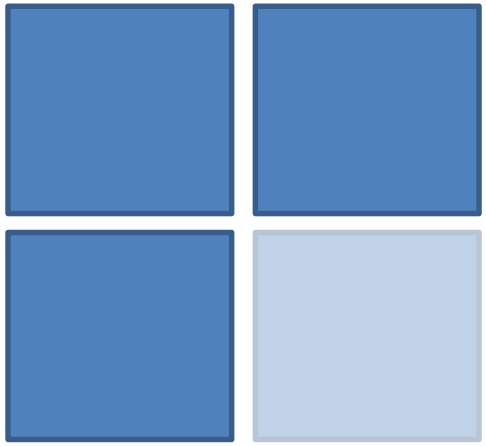
DnnWeaver achieves a speedup of up to **7.1x** Xeon.

Performance-per-watt over CPUs

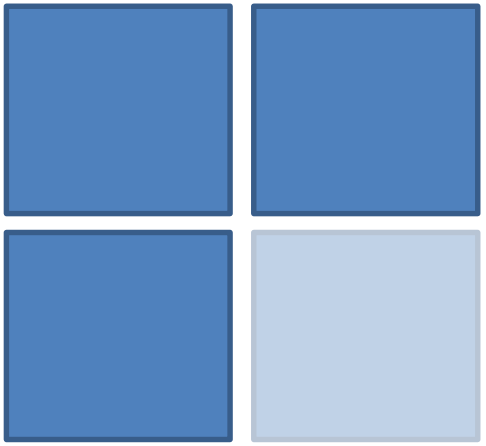


DnnWeaver is up to **36x** more power efficient than Xeon.

Conclusion

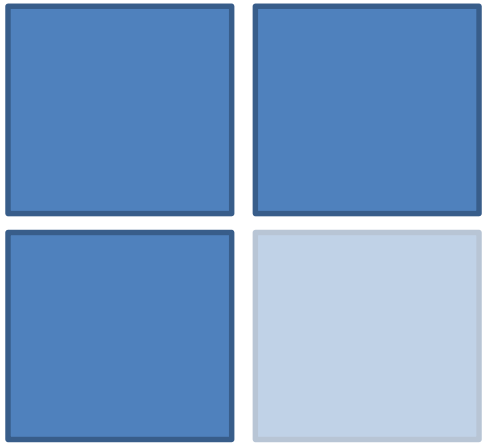


3 cores available



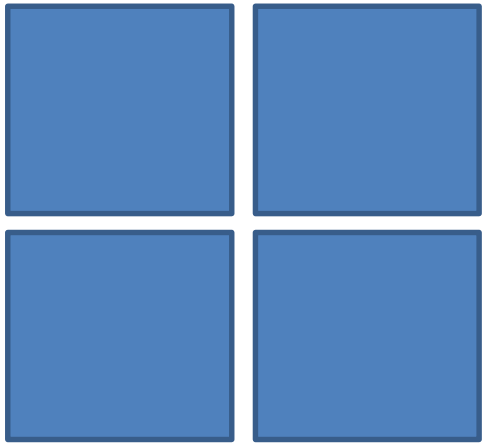
3 cores available

Conclusion



3 cores available

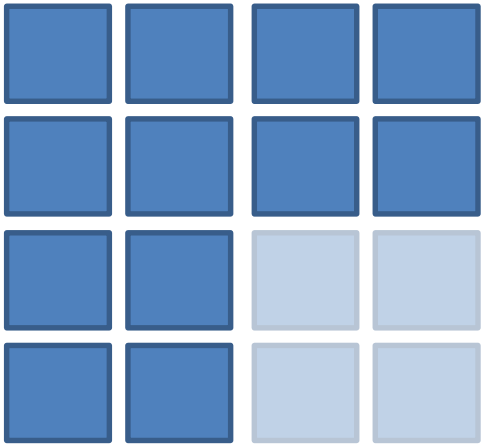
Microfluidic Cooling



4 cores available

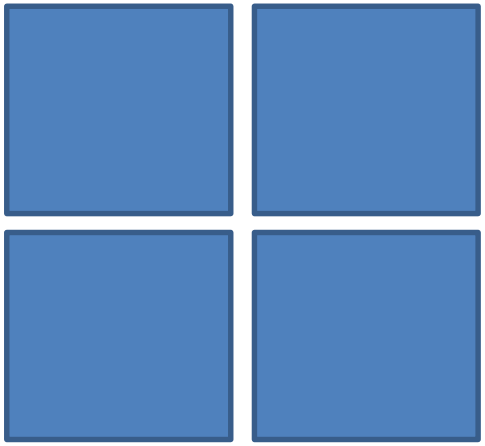
Conclusion

Approximate Computing



12 cores available

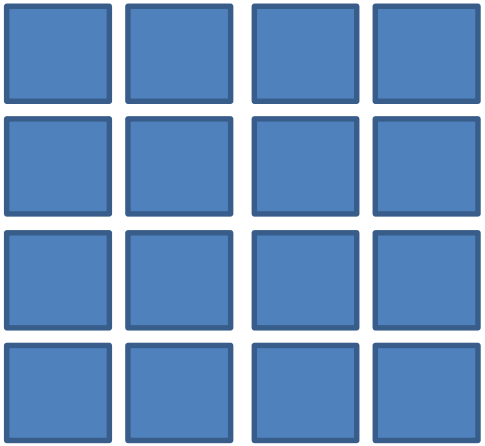
Microfluidic Cooling



4 cores available

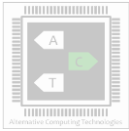
Conclusion

Approximate Computing
+
Microfluidic Cooling



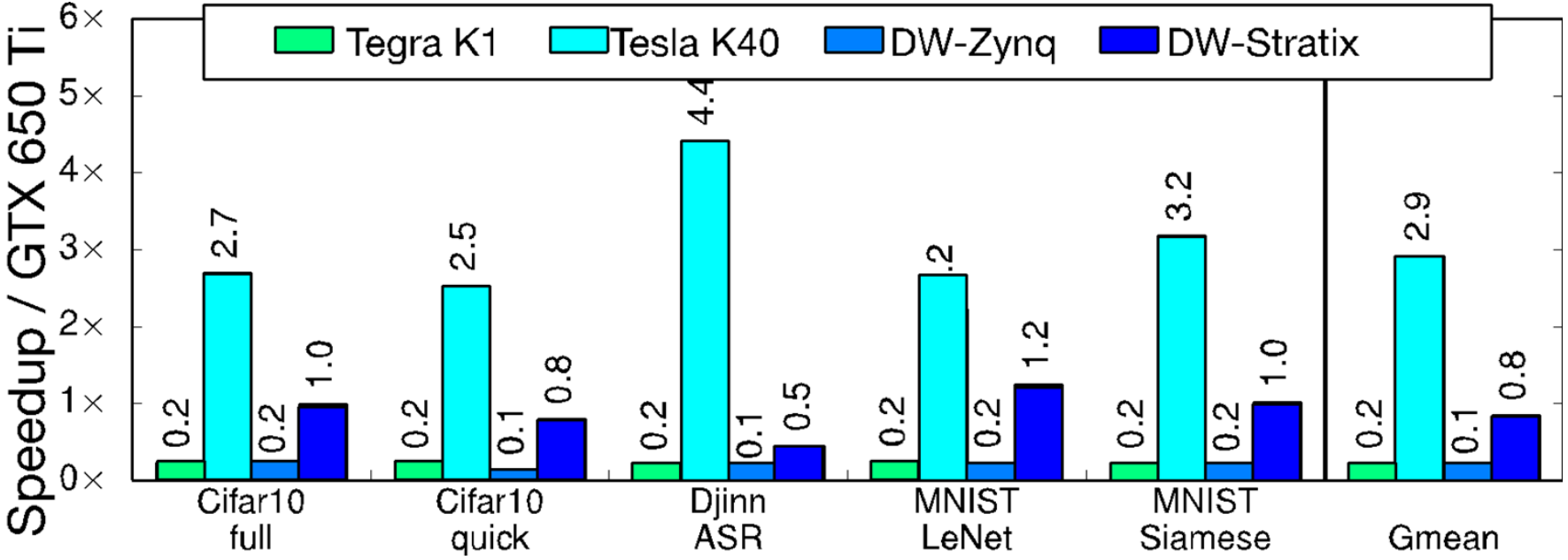
16 cores available

Questions?



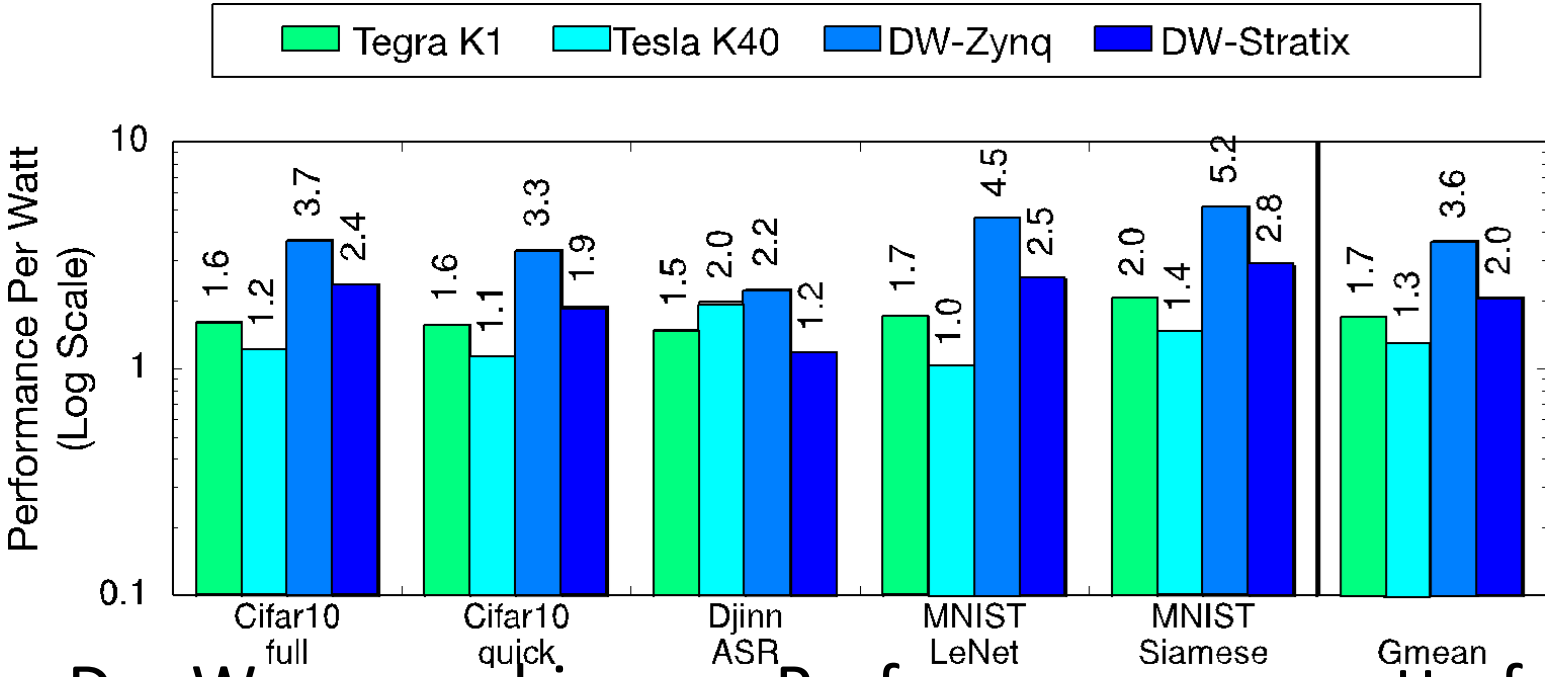
Backup

Speedup over GPUs



DnnWeaver provides an average of **0.8x** speedup over GTX 650Ti

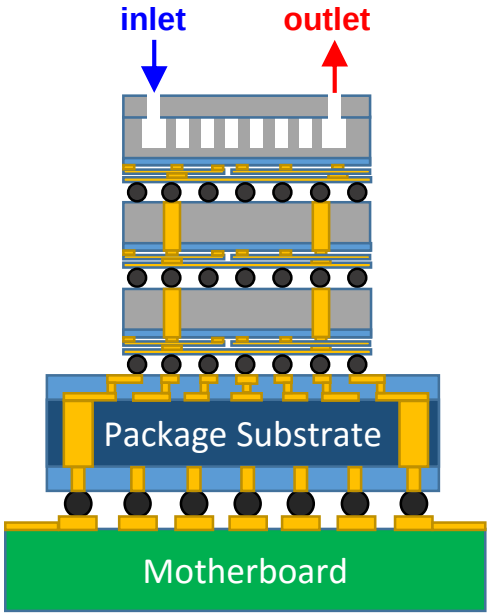
Performance-per-watt over GPUs

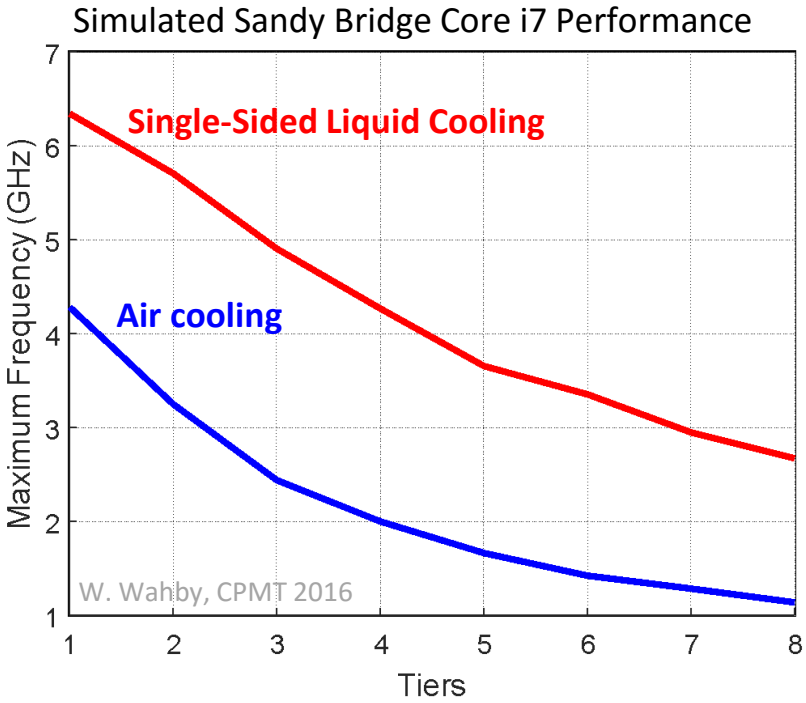


DnnWeaver achieves a Performance per watt of up to

3.6x GTX 650 Ti

3D stacking complicates heat removal





Tier-specific microfluidic cooling