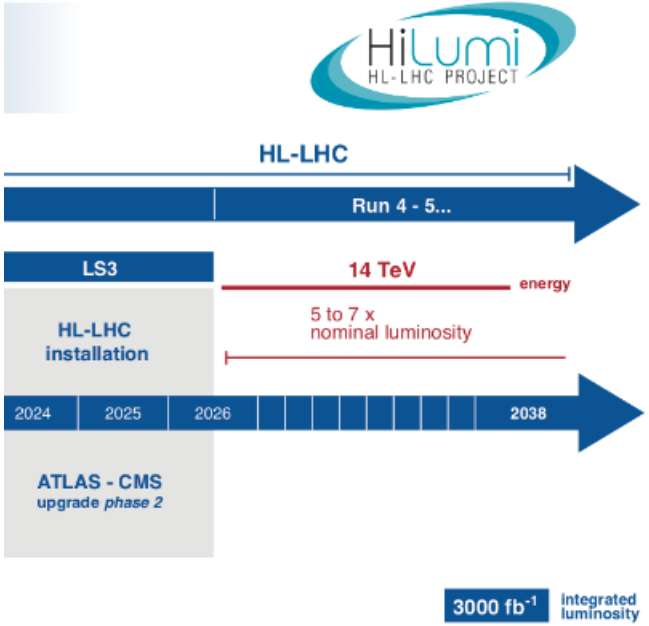
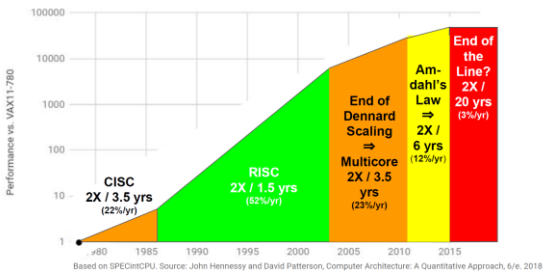
Computing Architectures in the HL-LHC Era

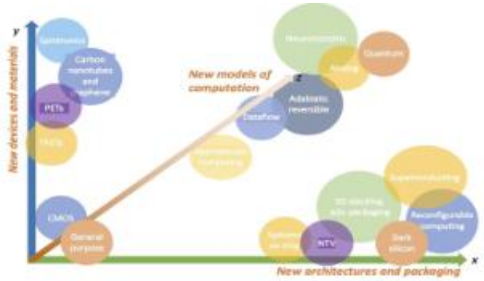
Paolo Calafiura (LBNL)

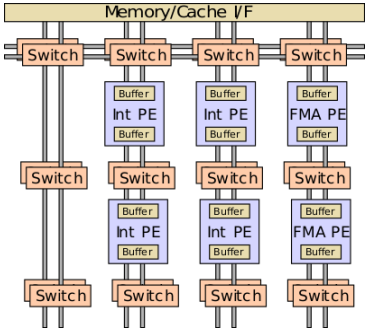
ACAT 2019, March 13



Overview/Disclaimers





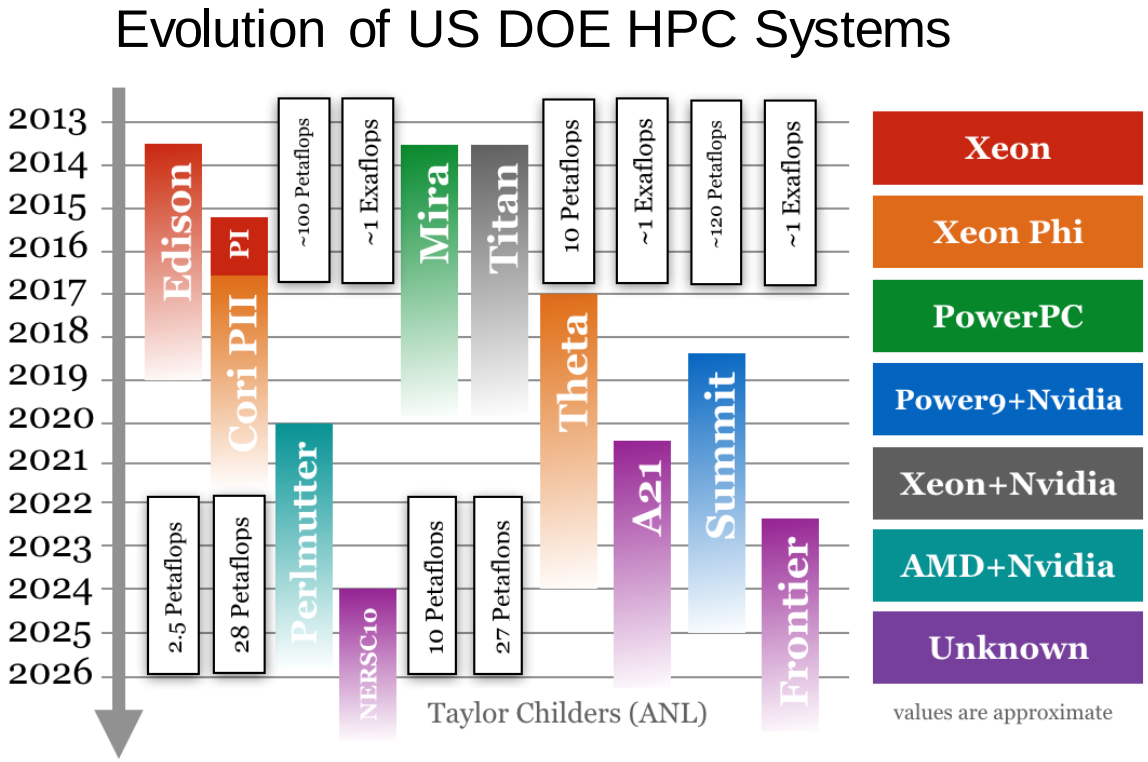


Apologies for the US and ATLAS bias

The only certainty about all HL-LHC predictions is that they will be proven wrong!

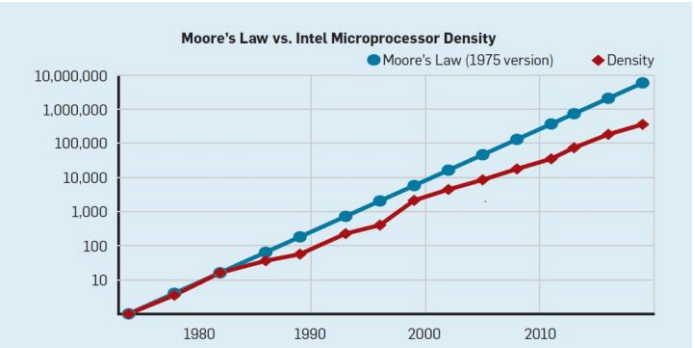
- but we can still learn something from them...

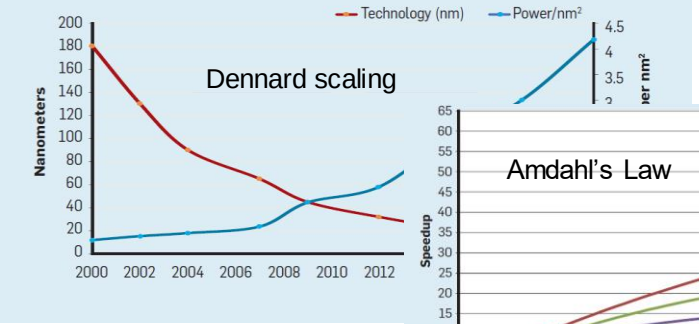
Rushing towards Heterogenous Computing

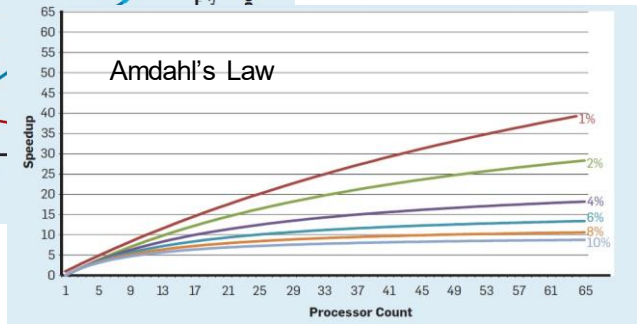


- **Over 20x more Flops available by 2026**
- **At most 10% will come from CPUs**
  - Multiple architectures within each system

Why is this happening?



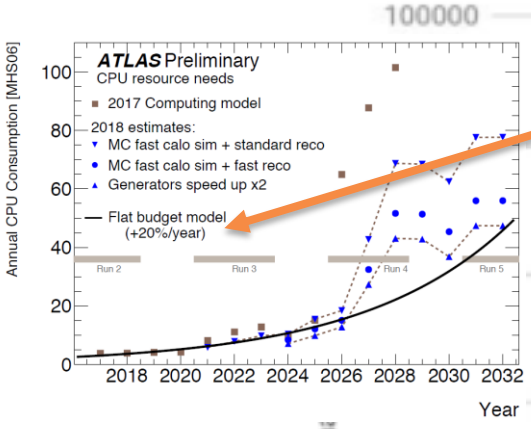




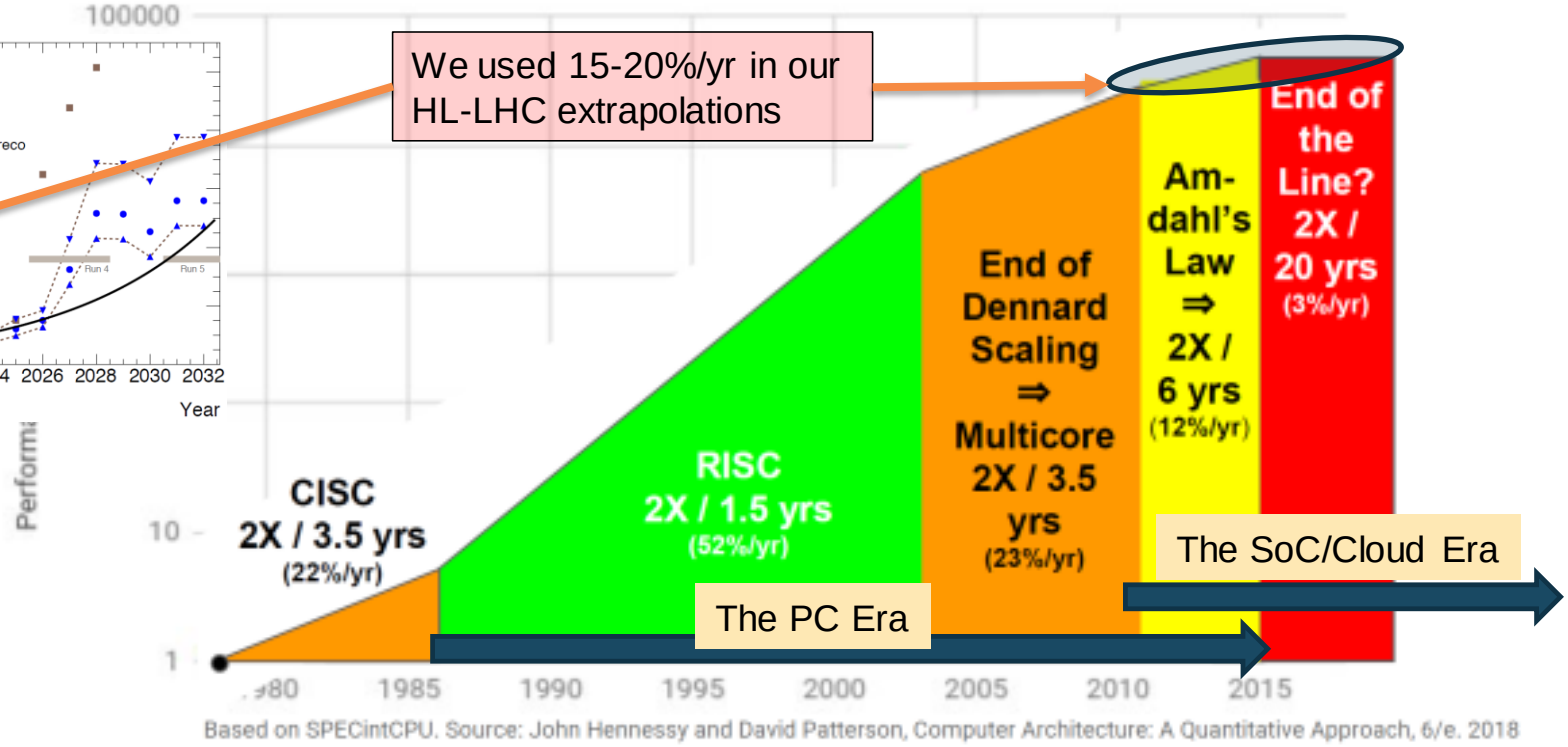
- **Transistor density keeps growing**
  - But 15x gap wrto 1975 Moore's law
- **Power keeps frequency down**
  - The multicore era
- **Amdahl limits core count**
  - HEP typically <~1%

Credit: John Hennessy, David Patterson

History of a Benchmark



We used 15-20%/yr in our HL-LHC extrapolations



Based on SPECintCPU. Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018

A New Golden Age for Computer Architecture

Credit:
Shalf, Leland

New Materials and Devices
The 2030s (10 year lead time)

Spintronics
Carbon nanotubes and graphene
PETs
TFETs
CMOS
General purpose

Felix@noon

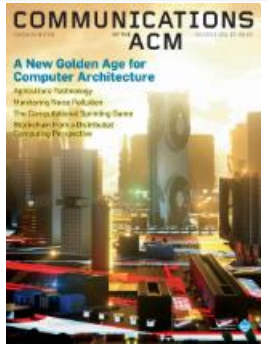
Heather yesterday,
Jean-Roch Monday

Neuromorphic
Quantum
Analog
New models of computation
Adiabatic reversible
Dataflow
Approximate computing
Superconducting
Domain Specific
3D stacking, adv. packaging
Reconfigurable computing
Systems on chip
NTV
Dark silicon

z

x

More Efficient Architectures and Packaging
The 2020s

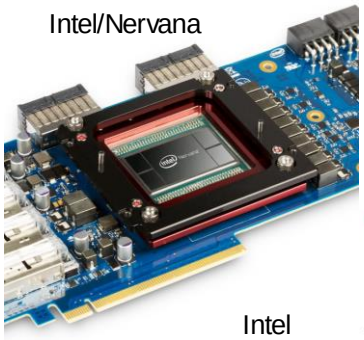


BERKELEY LAB

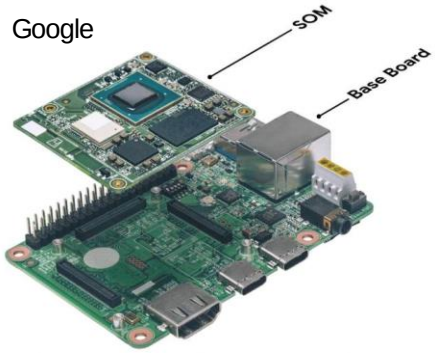
Domain Specific Architectures: Deep Learning

Source	NPU
Amazon	AWS Inferentia
Alibaba	Ali-NPU
Baidu	Kunlun
Bitmain	Sophon
Cambricon	MLU
Google	TPU
Graphcore	IPU
Groq	
Intel	NNP, Myriad, EyeQ
Nvidia	NVDLA
Huawei	Ascend
Apple	Neural Engine
Samsung	NPU
(wikichip)	

Intel/Nervana

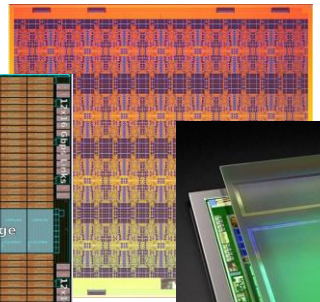


Google

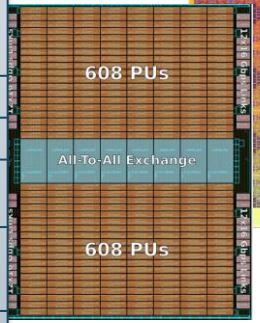


The Edge TPU development kit—SOM (above) and base board (below)

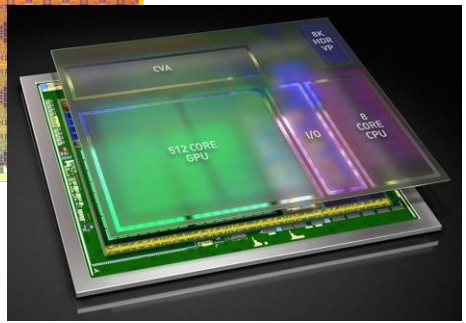
Intel

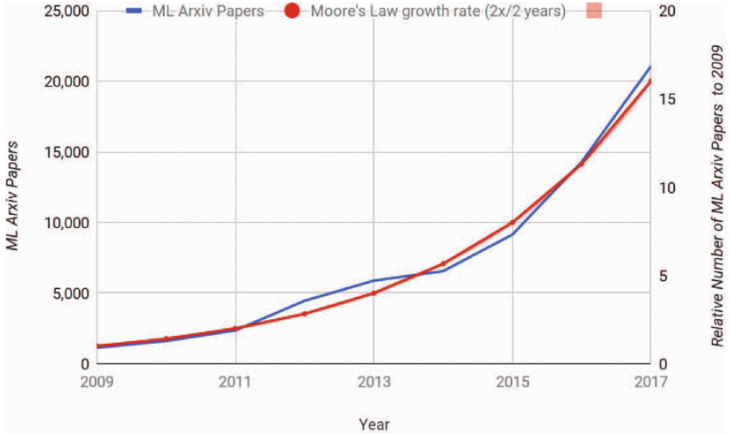


Graphcore



NVidia

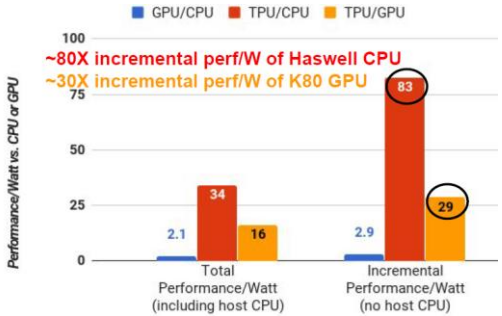


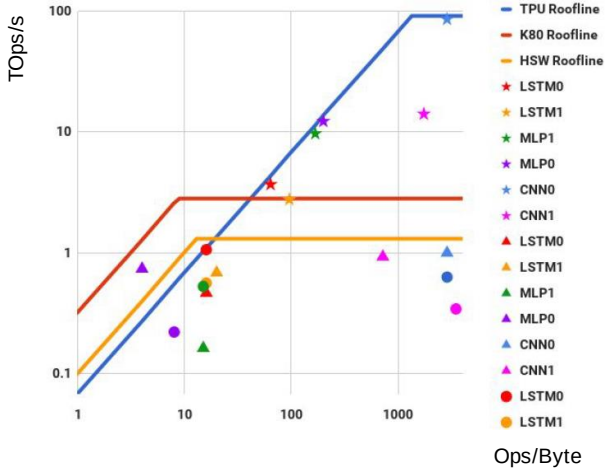


Credit: Dean, Patterson, Young

DSA: Neural Network Inference Examples

- **Google TPU**
  - Introduced to meet user inference needs (speech recognition)
  - First commercial implementation of a systolic array (matrix unit)
  - 50x more power efficient than a CPU
  - Edge TPU available as USB dongle



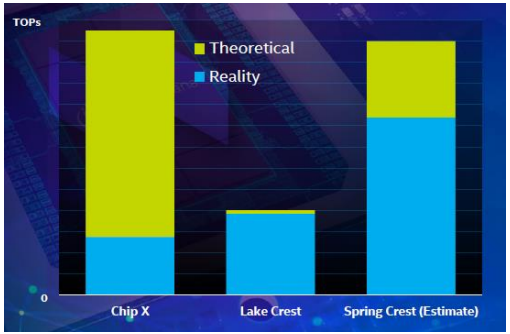


Source: Google

☆Star = TPU
△Triangle = GPU
○Circle = CPU

DSA: Neural Network Training Examples

- **Nvidia Volta Tensor Core**
  - 4-8x speedup wrto to Pascal
  - Mixed-precision matrix multiply
- **Intel/Nervana Lake Crest**
  - 32 GB 1 TB/s HBM2 Memory
    - 5x faster than GDDR5
  - Goal is to speed-up NN training 100x by 2020



Source: Intel

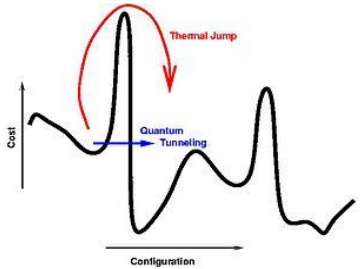
Domain Specific Architectures: QUBO

D-Wave Quantum Annealer

A special purpose computer that is designed to solve a particular optimization problem, namely finding the ground state of a classical Ising spin glass (Vazirany et al)

That it does (see Heather, Jean-Roch).
No consensus on whether:

- **D-Wave is a QA or a Simulated QA**
- **QA provides quantum speedup.**



wikipedia

D-Wave Instruction = $\sum_{i=1}^{N} a_i q_i + \sum_{i}^{N} \sum_{j}^{N} b_{ij} q_i q_j$

	2000Q (2017)	Pegasus (2020?)
# qubits (q_i)	2048	5640
# couplers (b_{ij})	6016	40484
Noise over D-Wave 1	~1/3	~1/10

Domain Specific Architectures: QUBO

Fujitsu Digital Annealer

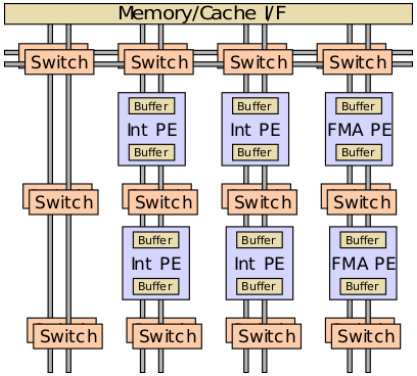
A next-generation architecture inspired by quantum phenomena, for the high-speed resolution of combinatorial optimization problems. (Fujitsu)

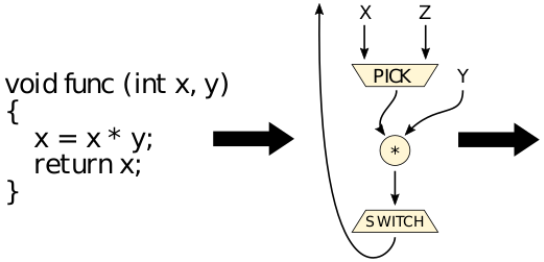
- **Qubits replaced by "bit updating blocks" with on-chip memory for its a_i and b_{ij}s**
- **"Logic blocks" perform bit flips**
- **In principle no embedding and noise problems**
- **1M blocks annealer under development**
- **Looking forward to test it!**

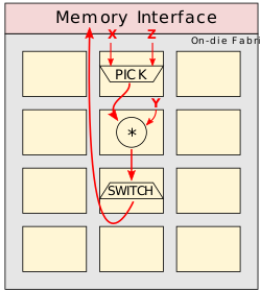
DAU Instruction = $$\sum_{i=1}^{N} a_i q_i + \sum_{i}^{N} \sum_{j}^{N} b_{ij} q_i q_j$$

	DAU 1 (2018)	DAU 2 (2019)
# blocks (q_i)	1024	8192
# couplers (b_{ij})	1024*1024	8192*8192
Precision b_{ij}	16 bits	max 64 bits

New Computation Models: Dataflow Engines



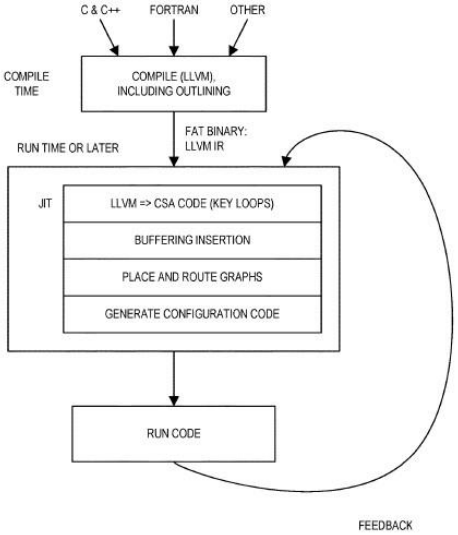




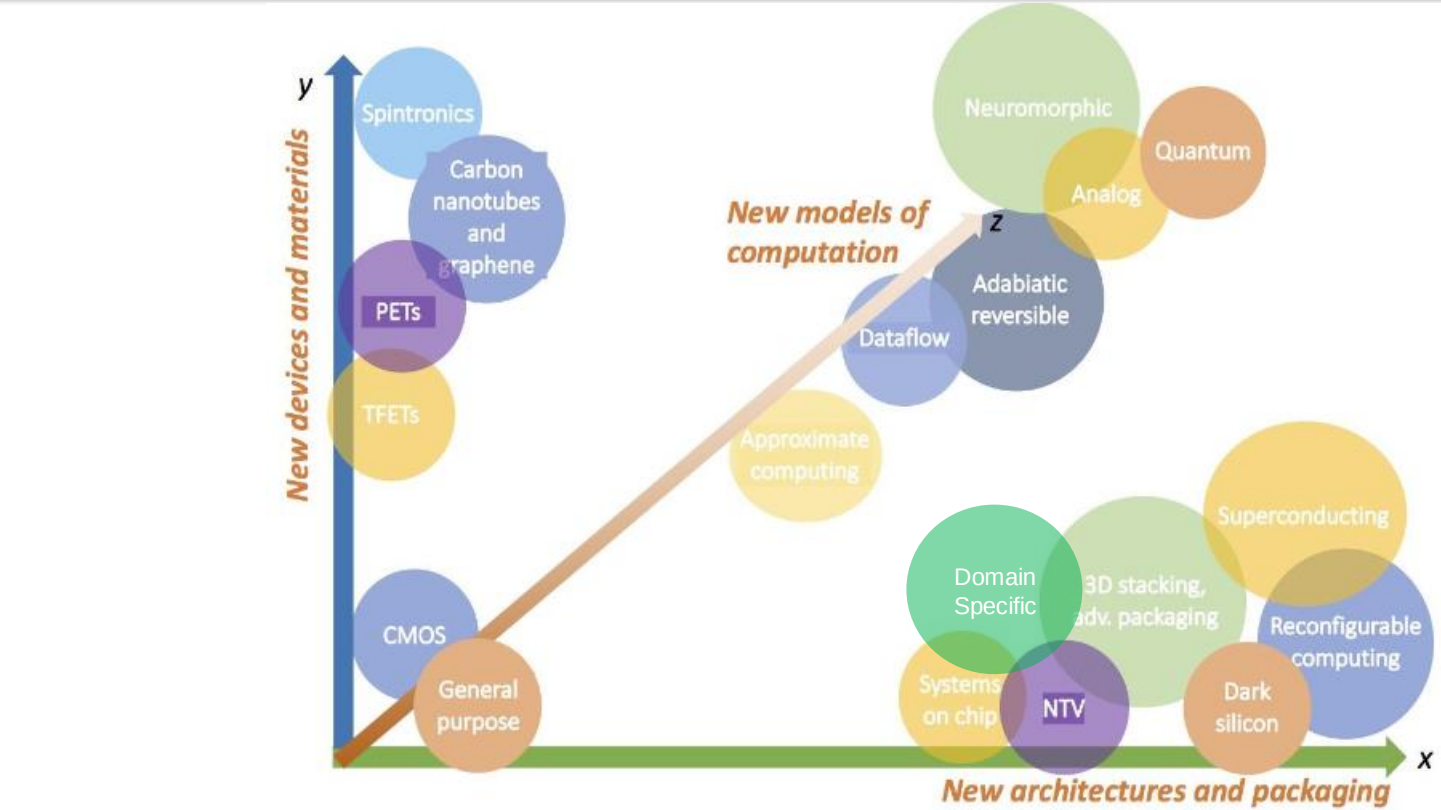
Sources:
wikichip,
nextplatform

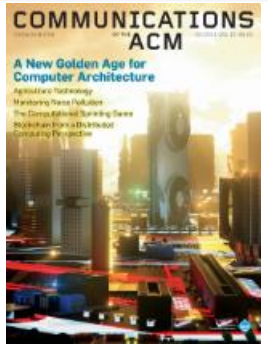
Intel Configurable Spatial Accelerator

- Dataflow graph engine
  - Maps compiler execution graph (IR) to hardware fabric
- Elements of modern switches, FPGAs, KNL
- Rumored to be architecture of ANL Aurora A21



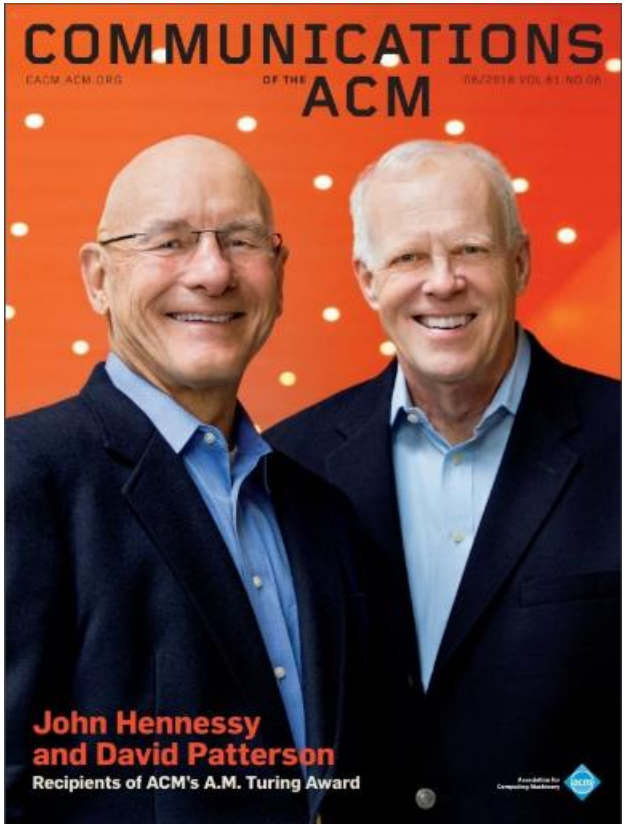
A New Golden Age for Computer Architecture?





Many unproven candidates yet to be investigated at scale. Most are disruptive to our current ecosystem.

A New Golden Age for Computer Architects!



Developers left to deal with the uncertainty

The Dharmas of HL-LHC Software

Coding for the Unknown

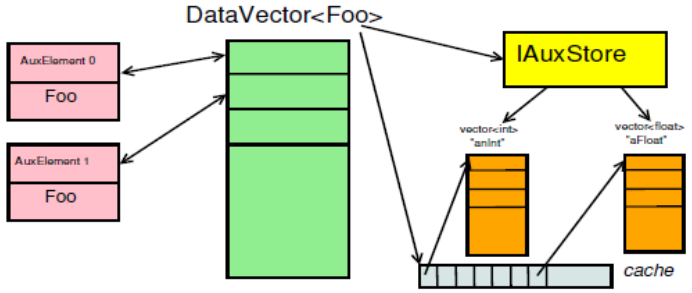
- **Automated performance portability is an old dream**
  - A GPU does not work like a TPU, much less like a CPU
  - No magic C++ compiler or library yet
  - Domain Specific Languages (e.g. MatLab, TensorFlow) may help capture deep, universal abstractions
    - to some extent, so may libraries like eigen
    - and frameworks like SYCL, Kokkos, and Alpaka (Michael talk Monday)

Coding for the Unknown

- **Data Model portability is a must**
  - Moving, copying or reformatting data kills performance
    - E.g. data conversion 40% of mkFit running time (Mario Monday)
      - ATLAS HLT had similar results (15-20% to conversion)
- **To first approximation, the only portable performant data structures are arrays of basic types**

A First Step: ATLAS DataVector

- ***Extensible*** **SoA container that looks and feels like a polymorphic AoS container (ref)**
- **Improves memory locality, usability, I/O performance**
  - Still relies on ROOT for pointer swizzling



From Events to Frames to Awkard Arrays?

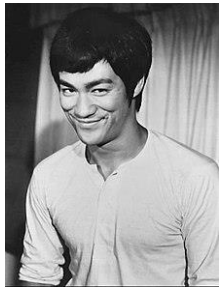
- **Event processing not a good match to accelerators**
- **Need a container-centric programming model**
  - Don’t call us we will call you:
    - Lambda functions operating on objects from multiple containers in multiple events
      - Complication: HEP Data objects are nested, linked, irregular

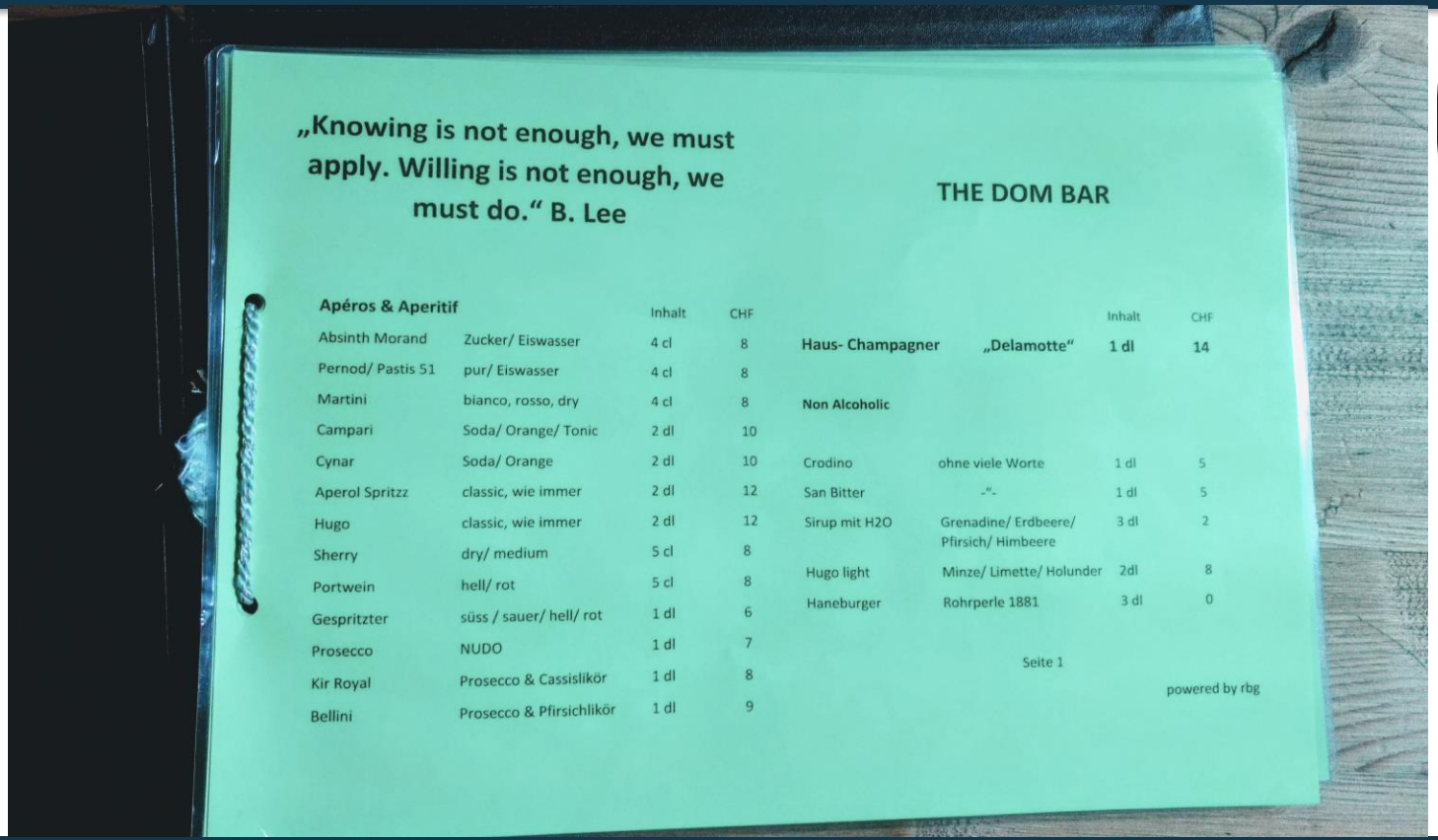
Looking forward to Jim P talk on Thu

Final Thoughts

- **We don’t really have to abandon our C++, x86, GRID computing model**
  - But the “New Golden Age” of hardware specialization is the proverbial crisis we should not let go to waste
- **The revolution starts from the Data Model**
  - And continues with container-centric programming

Final Final Thought: The Dom Sutta





„Knowing is not enough, we must apply. Willing is not enough, we must do.“ B. Lee

THE DOM BAR

Apéros & Aperitif		Inhalt	CHF
Absinth Morand	Zucker/ Eiswasser	4 cl	8
Pernod/ Pastis 51	pur/ Eiswasser	4 cl	8
Martini	bianco, rosso, dry	4 cl	8
Campari	Soda/ Orange/ Tonic	2 dl	10
Cynar	Soda/ Orange	2 dl	10
Aperol Spritzz	classic, wie immer	2 dl	12
Hugo	classic, wie immer	2 dl	12
Sherry	dry/ medium	5 cl	8
Portwein	hell/ rot	5 cl	8
Gespritzter	süss / sauer/ hell/ rot	1 dl	6
Prosecco	NUDO	1 dl	7
Kir Royal	Prosecco & Cassislikör	1 dl	8
Bellini	Prosecco & Pfirsichlikör	1 dl	9

		Inhalt	CHF
Haus- Champagner	„Delamotte“	1 dl	14
Non Alcoholic			
Crodino	ohne viele Worte	1 dl	5
San Bitter	-“-	1 dl	5
Sirup mit H2O	Grenadine/ Erdbeere/ Pfirsich/ Himbeere	3 dl	2
Hugo light	Minze/ Limette/ Holunder	2dl	8
Haneburger	Rohrperle 1881	3 dl	0

Seite 1

powered by rbg

Thanks!

John Shalf, Charles Leggett, Illya Shapoval, Vakho Tsulaia, Jim Pivarski

AIHENP 96 was my first computing conference. Nice to be back!