Low-Entropy Computing Systems

A direction of energy efficient computing

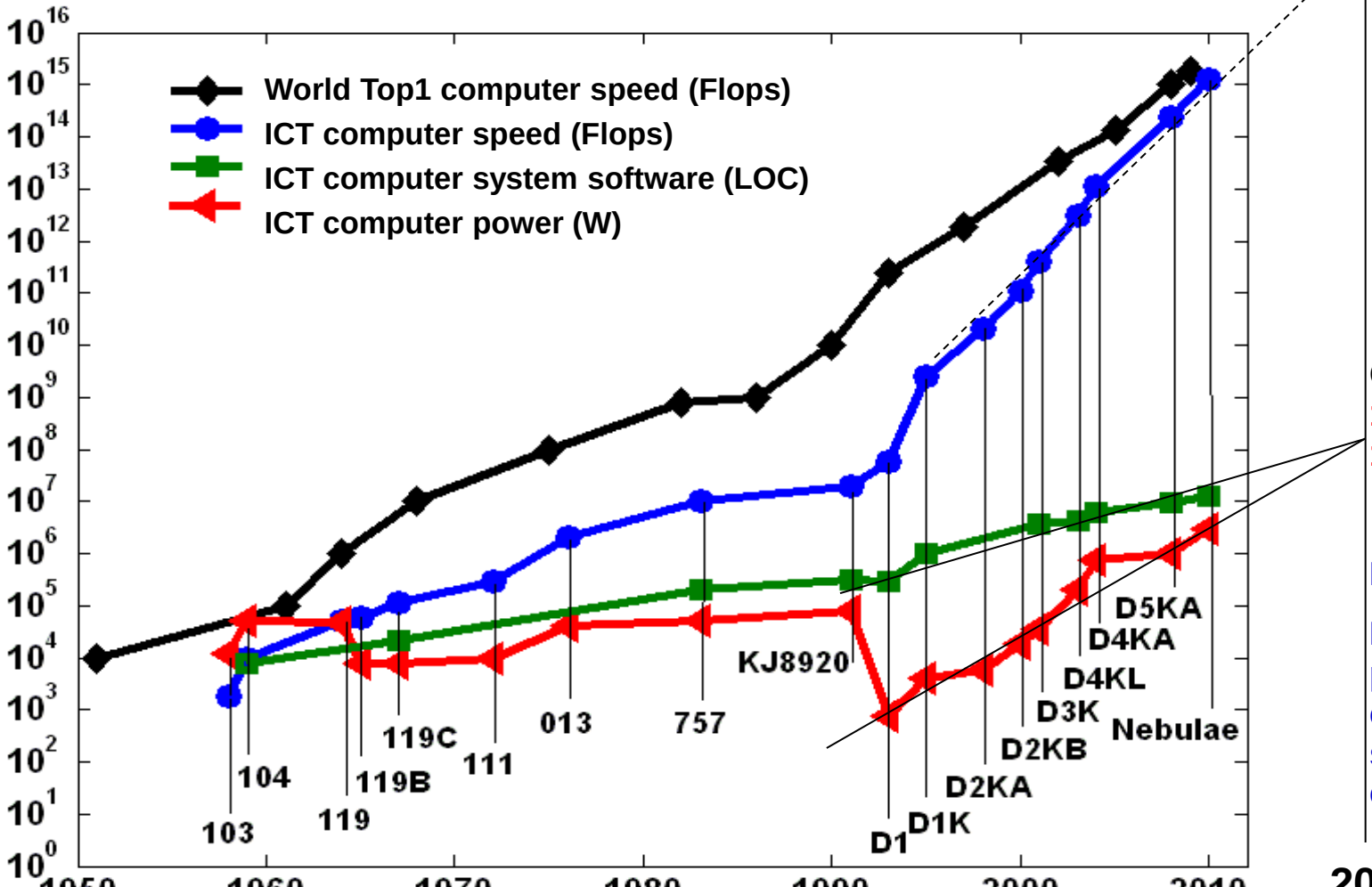
Zhiwei Xu
Institute of Computing Technology (ICT)
Chinese Academy of Sciences (CAS)
http://novel.ict.ac.cn/zxu/
zxu@ict.ac.cn

Research supported in part by the CAS Strategic Priority Program (XDB02040009), the NSF of China (61532016), and the MOST (2016YFB1000200).

Trends of ICT Developed Computers

Speed, software complexity, power

Z Xu, G Li, Computing for the Masses,
***Communications of ACM*, 54(10): 129-137 (2011)**



Exaflops (10^{18})
Datacenter for
100's M (10^{8}) users

Often overlooked!
100 M (10^{8}) LOC
100 M (10^{8}) W

Needs:

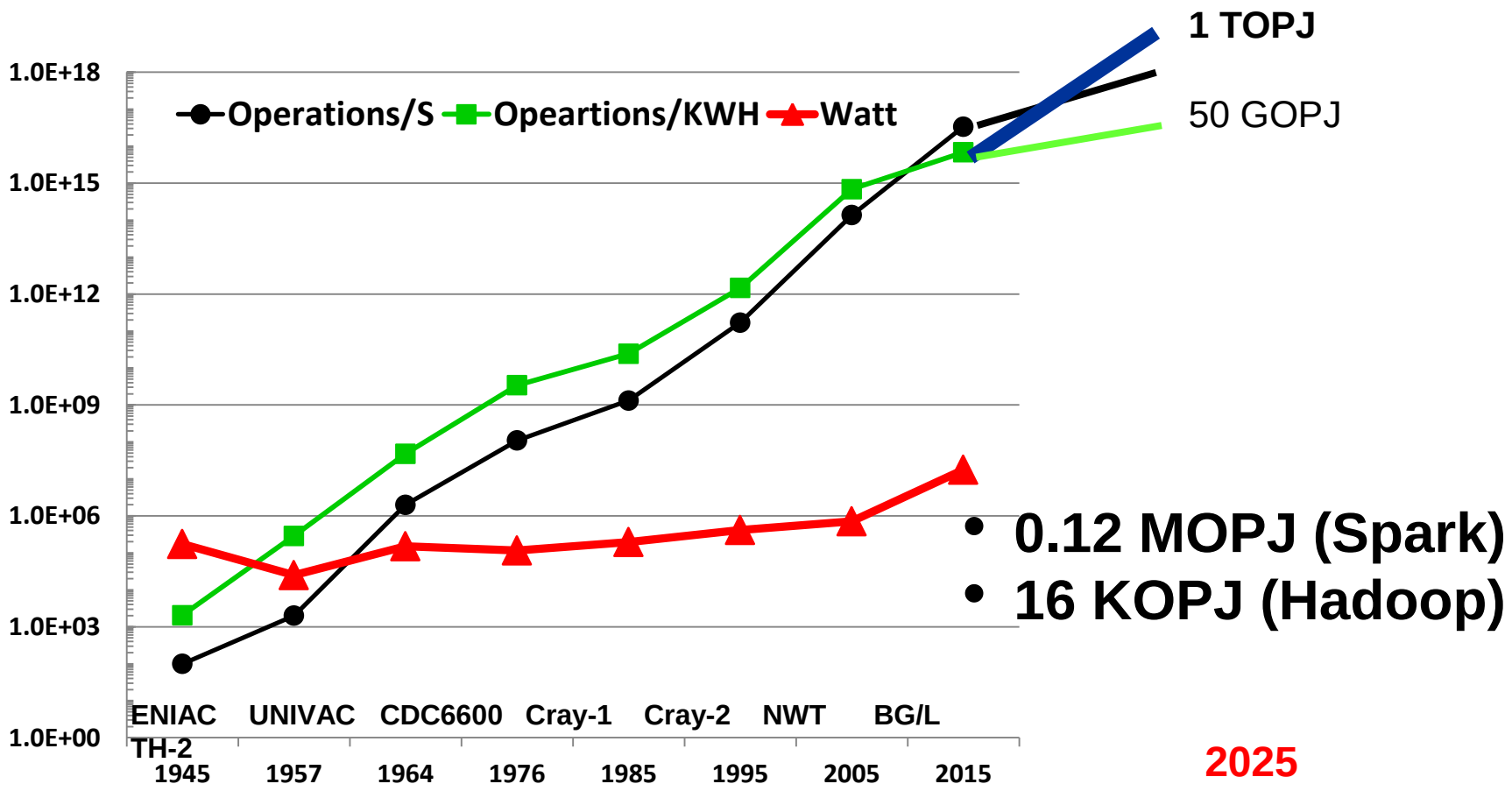
Maintain growth in performance, but control power & system software complexity

2020-2030

Fundamental Challenge

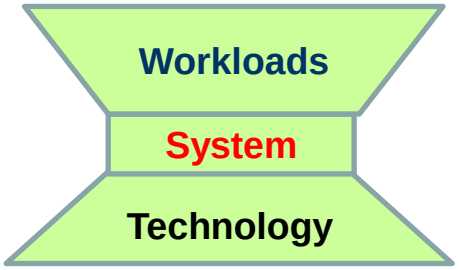
First time in 70 years

- Energy efficiency growth lags behind speed growth
  - Big data computing is especially bad: **15-150 KOPJ**



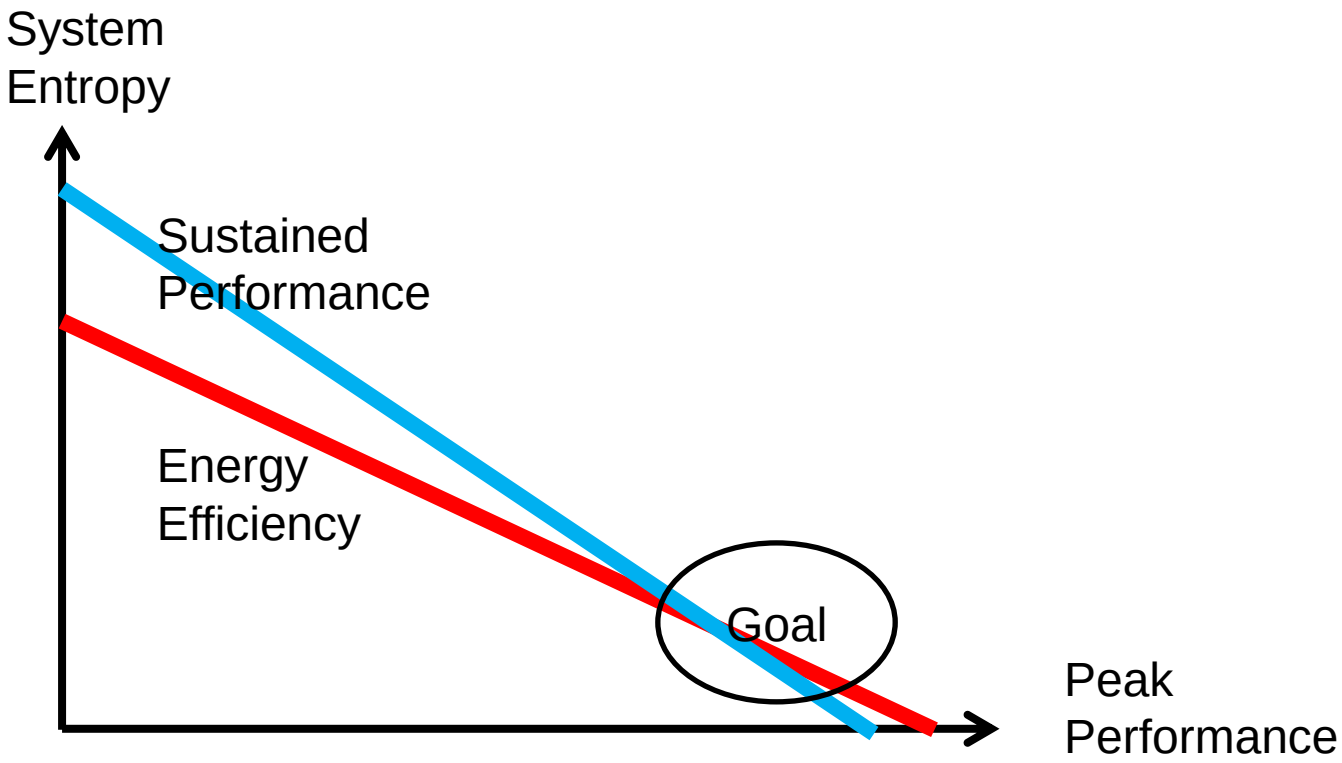
Workload Complexity and Sys Disorder Keep Growing

- Three types of workloads share a Sugon HPC system
  - Traditional HPC (scientific and engineering)
  - Data analytics (online and offline)
  - Machine learning, especially deep learning
- Many types of disorder exist
  - Workload dynamicity and uncertainty
  - Interferences of workloads
  - System jitter (clutter, noise)
  - Impedance mismatch
  - **Unbounded flexibility** (Gordon Bell: general-purpose always wins)
- Efficiency needs order
  - Uncertainty bounding will be a main, fundamental challenge
    - A key metric: computing systems entropy (cf., tail latency)
    - Low-entropy systems with disciplined flexibility (Symphony vs. Bazaar)



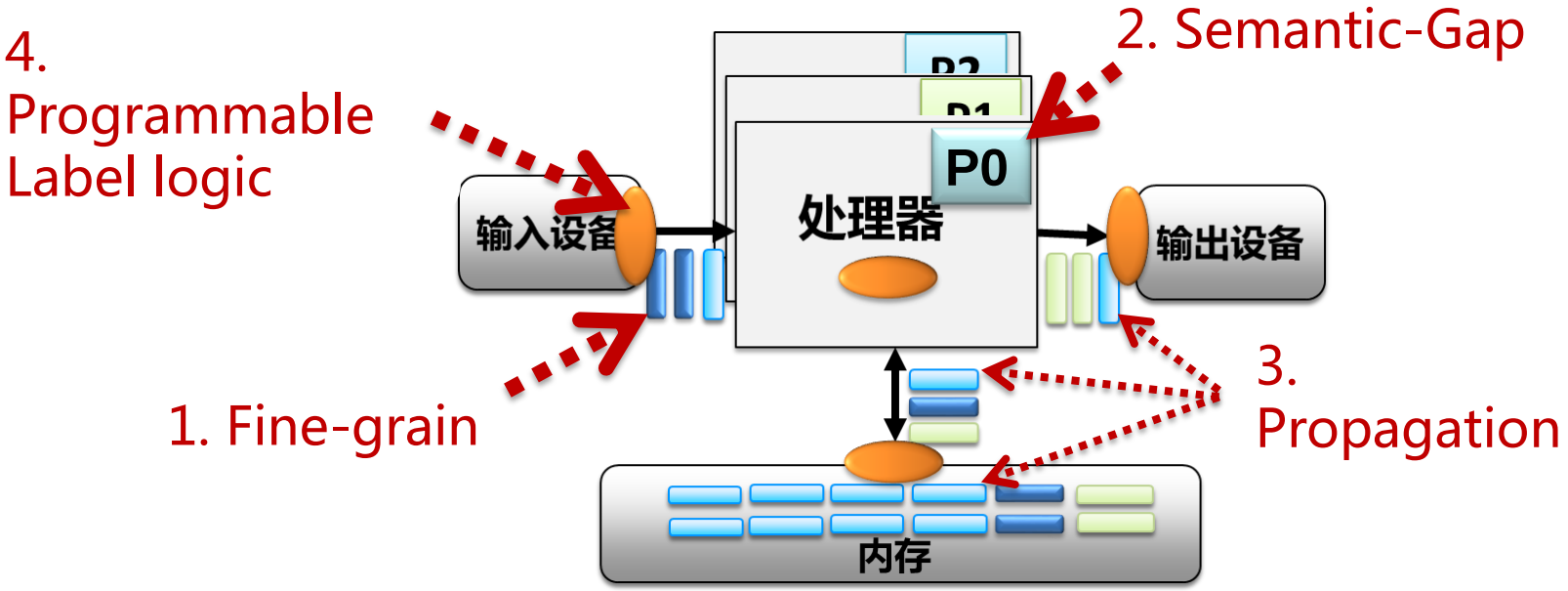
Low Entropy Hypothesis

- As computing system entropy decreases,
  - Energy efficiency increases, and
  - Sustained performance increases



Labeled von Neumann Architecture

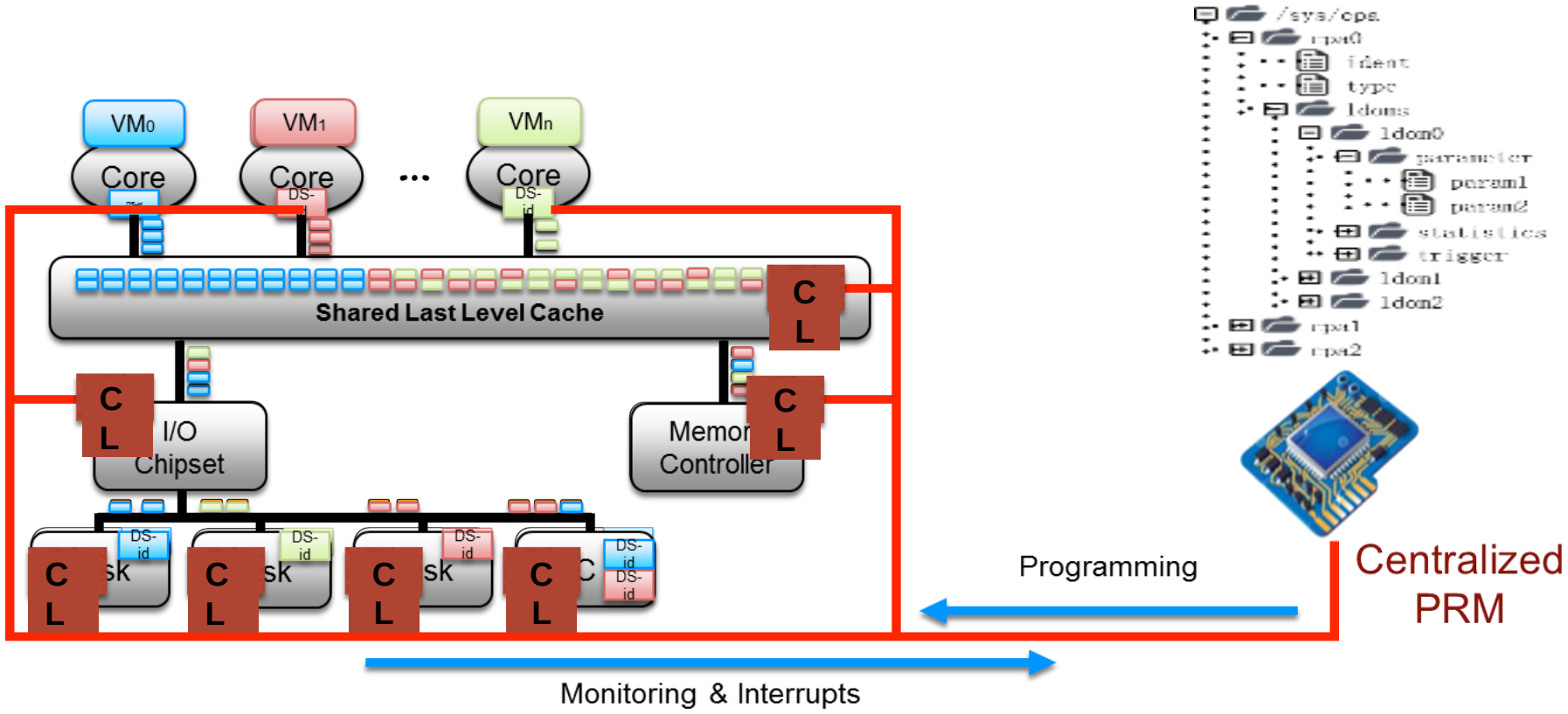
- **Fine-grain**: attach a label to each memory and I/O request
- **Semantic-Gap**: relate labels to VM/Proc/Thread/Var
- **Propagation**: propagate labels in the whole machine
- **Programmable label logic**: provide differentiated services based on different label-indexed rules



Bao and Wang, Labeled von Neumann Architecture for Software-Defined Cloud, Journal of Computer Science and Technology, 2017 Vol. 32 (2): 219-223.

Programmable Architecture for Resourcing-on-Demand

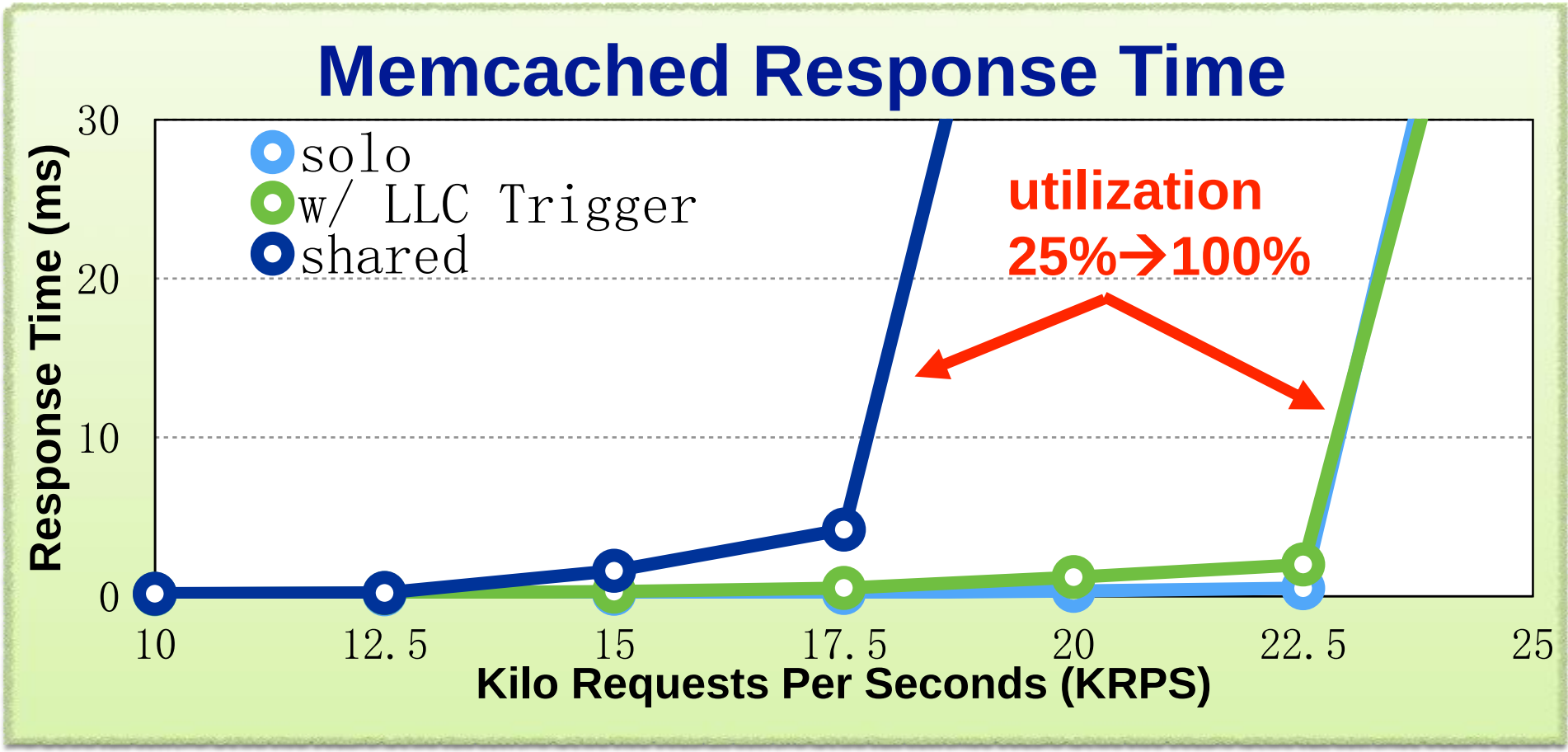
Ma et. al, Supporting Differentiated Services in Computers via Programmable Architecture for Resourcing-on-Demand (PARD), ***ASPLOS**, 2015*



Improve Utilization w/o Loss of QoS

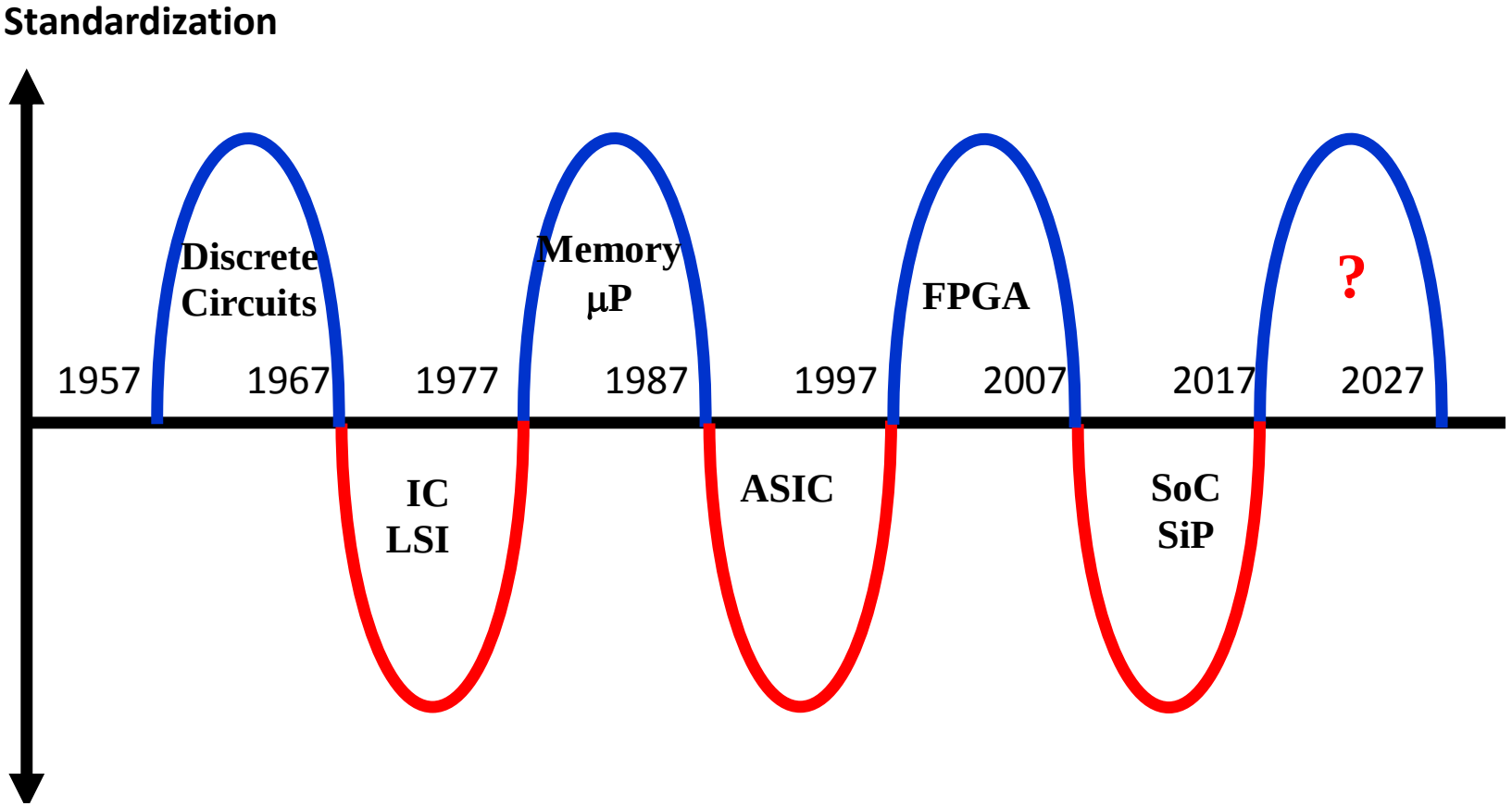
CPU Utilization 4X

- Memcached： Tail Latency <1.5ms



Makimoto's Wave

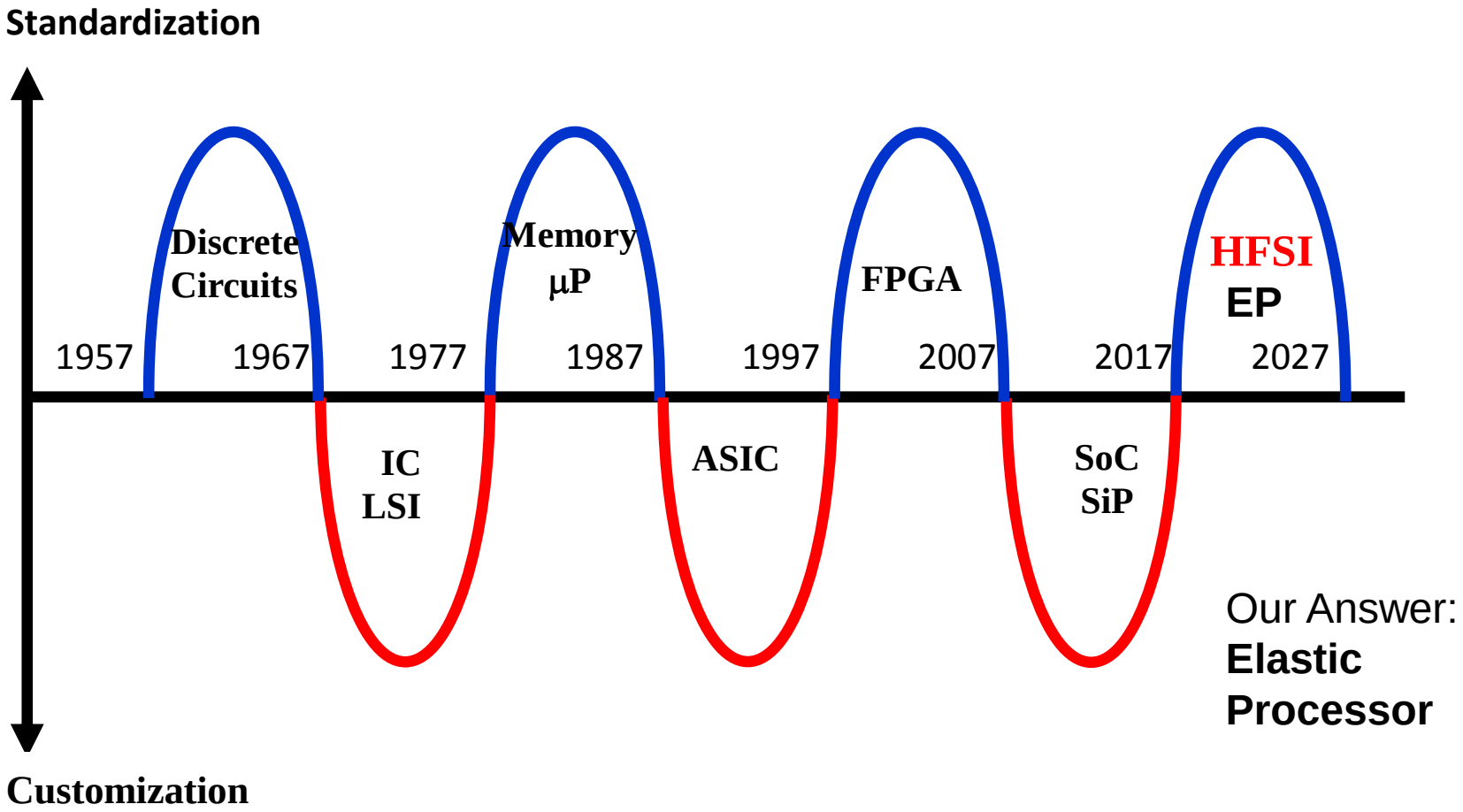
- Semiconductor technology will soon enter another phase change. But what is it?



Tsugio Makimoto, "Implications of Makimoto's Wave", *IEEE Computer*, vol. 46, no. , pp. 32-37, Dec. 2013

Makimoto's Wave

- HFSI: Highly Flexible Super Integration
  - Redundant circuits can be shut off when not in use



Our Answer:
Elastic Processor

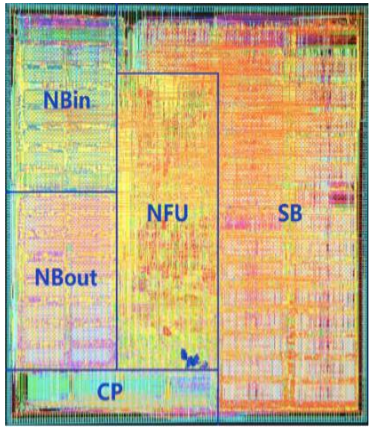
Elastic Processor

- A new architecture style (FISC)
  - Featuring **function instructions** executed by **programmable ASIC** accelerators
  - Targeting 1000 GOPS/W = 1 TOPJ

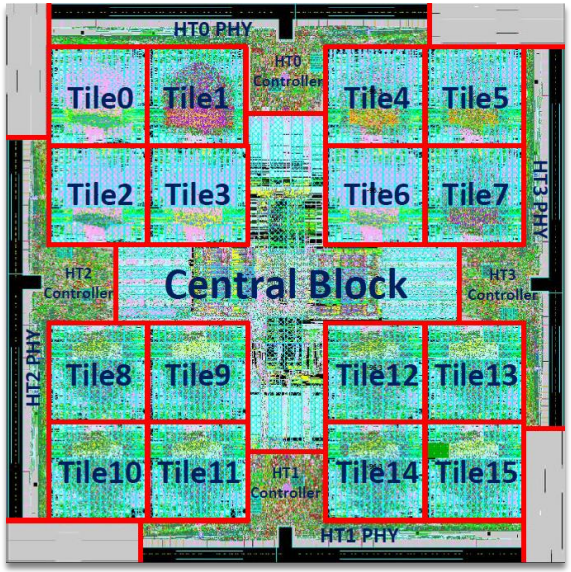
	CISC Intel X86	RISC ARM	FISC Function Instruction Set Computer
Chip types:	**10s**	**1K**	**10K**
Power:	**10~100W**	**1~10W**	**0.1~1W**
Apps/chip:	**10M**	**100K**	**10K**

The DianNao Family Showing Potential for TOPJ

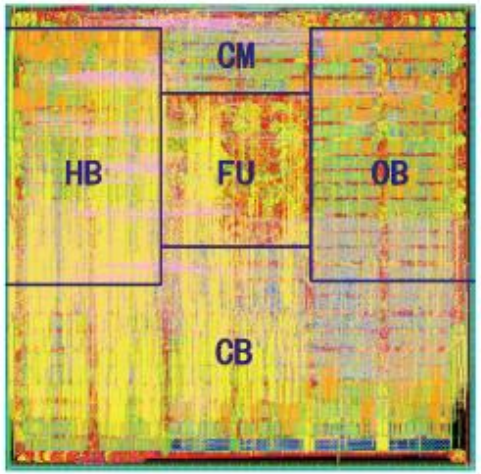
Yunji Chen, Tianshi Chen, Zhiwei Xu, Ninghui Sun, Olivier Temam. DianNao family: energy-efficient hardware accelerators for machine learning. *Communications of the ACM* 59(11): 105-112 (2016) Research Highlights paper



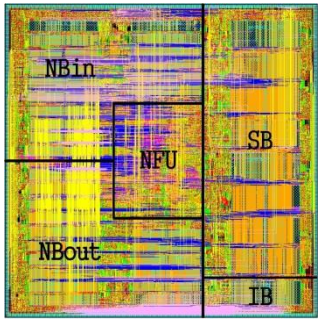
“电脑”
DianNao
931 GOPS/W
ASPLOS’14 Best Paper



“大电脑”
DaDianNao
100-250 GOPS/W
MICRO’14 Best Paper



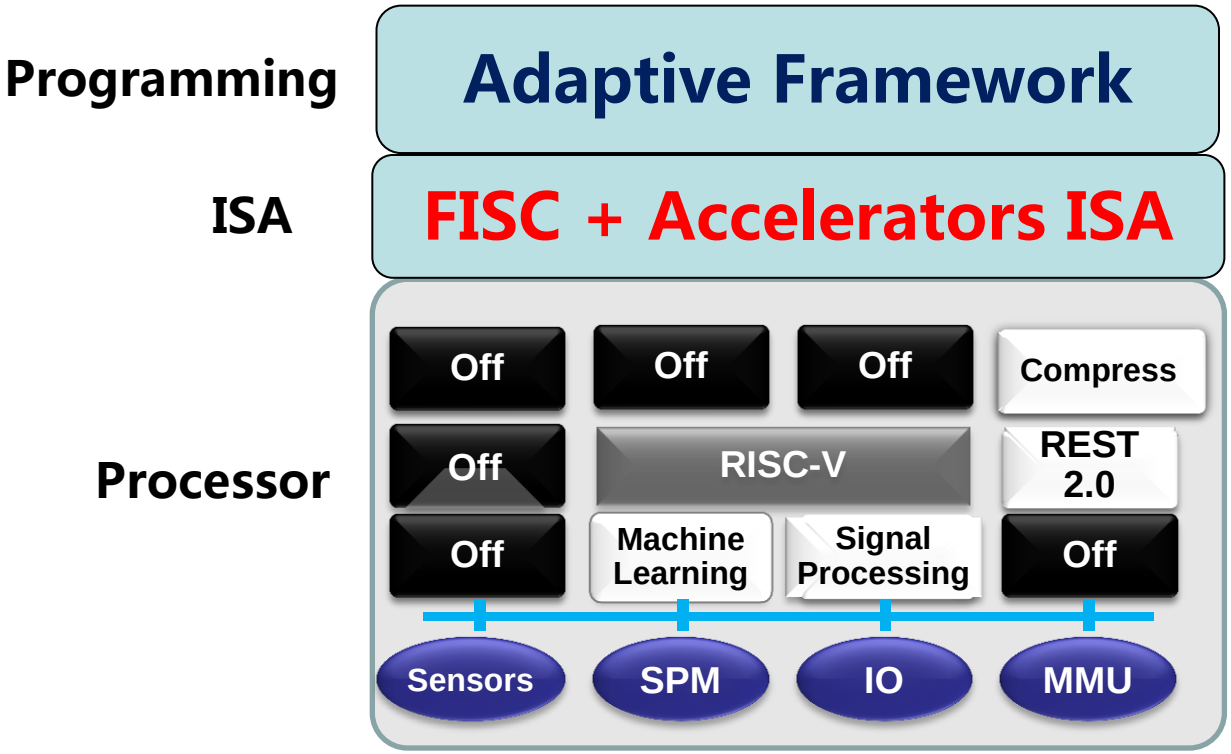
“普电脑”
PuDianNao
300-1200 GOPS/W
ASPLOS’15



“视电脑”
ShiDianNao
2-4 TOPS/W
ISCA’15

Elastic Processor

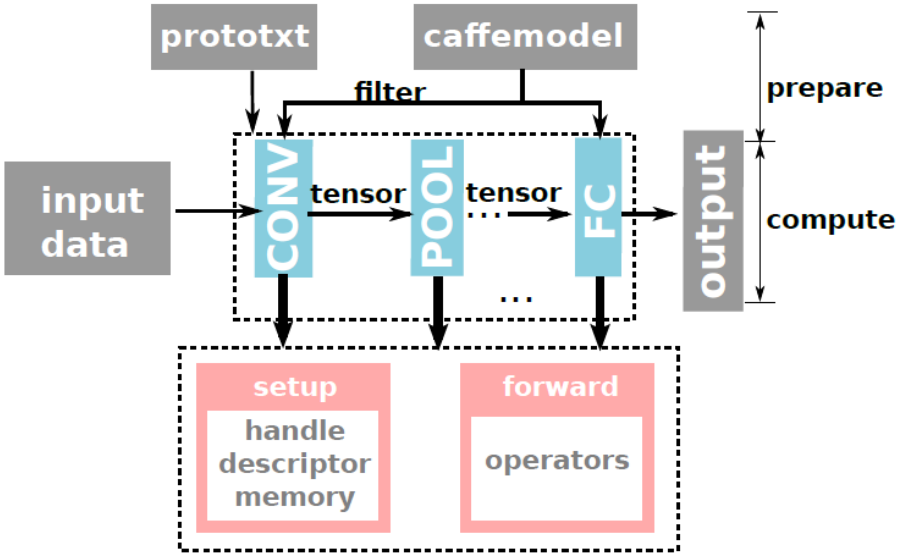
- Accelerators dominate (>99% of the time)
  - Dynamic customization, low switching overhead
  - Accelerator ISA
- Adaptive runtime with >50% efficiency

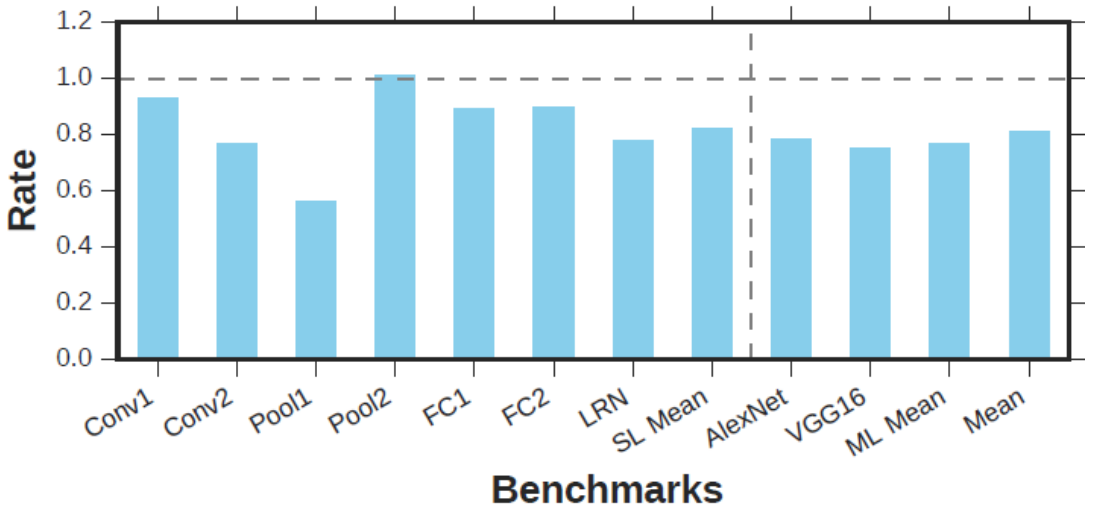


Liu et al, Cambricon: An Instruction Set Architecture for Neural Networks, in *Proceedings of the 43rd ACM/IEEE International Symposium on Computer Architecture (ISCA'16)*, 2016.

Adaptive Framework
DLPLib

- No pollution to Caffe
- Tensor+Filter
- 79% of assembly codes performance





Lan et al, DLPlib: A Library for Deep Learning Processor, Journal of Computer Science and Technology, 2017 Vol. 32 (2): 286-296.

zxu@ict.ac.cn http://novel.ict.ac.cn/zxu/