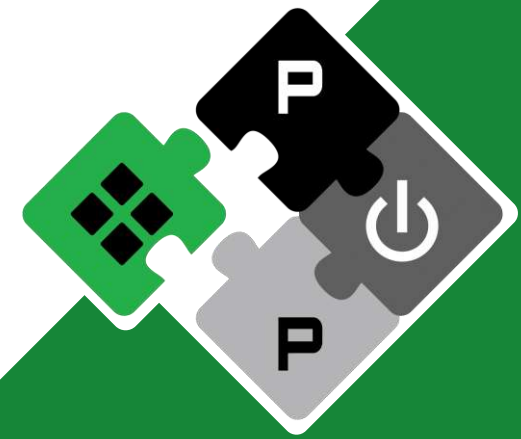
PULP: Looking Back and Looking Forward

Luca Benini lbenini@ethz.ch, luca.Benini@unibo.it



PULP Platform

Open Source Hardware, the way it should be!

@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

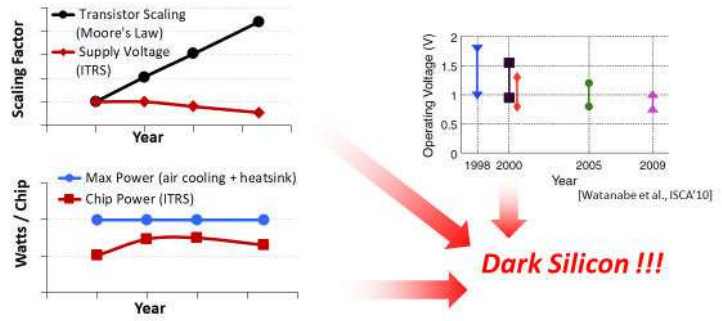
Looking Back: April 2012, Job Talk @ ETHZ

life.augmented

Digital Platform Design in the Twilight of Moore's Law

Luca Benini
Università di Bologna & STMicroelectronics

The Twilight of Moore's Law: Power



Thermal wall: transistor count still increases exponentially but we can no longer power the entire chip (voltages, cooling do not scale)

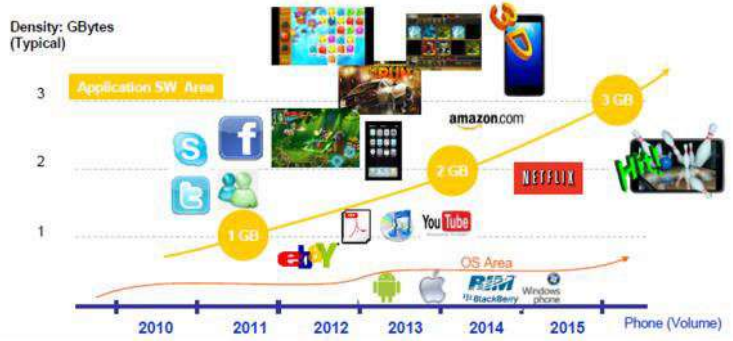
[Hardavellas11]

The Twilight of Moore's Law: Economics

130nm	90nm	65nm/55nm	45/40nm	32/28nm	22/20nm
Altis Semiconductor	Dongbu Hitek	Freescale	Fujitsu	Globalfoundries	Globalfoundries
Dongbu Hitek	Freescale	Fujitsu	Panasonic	Intel	Intel
Freescale	Fujitsu	Globalfoundries	Globalfoundries	Panasonic	Samsung
Fujitsu	Globalfoundries	IBM	IBM	Samsung	ST Microelectronics
Globalfoundries	Grace Semiconductor	Infineon	Intel	ST Microelectronics	TSMC
Grace Semiconductor	IBM	Intel	Renesas (NEC)	TSMC	UMC
IBM	Infineon	Panasonic	Samsung	UMC	
Infineon	Intel	Renesas (NEC)	SMIC		
Intel	Panasonic	Samsung	Sony		
Panasonic	Renesas (NEC)	SMIC	ST Microelectronics		
Renesas (NEC)	Samsung	Sony	Texas Instruments		
Samsung	Seiko Epson	ST Microelectronics	Toshiba		
Seiko Epson	SMIC	Texas Instruments	TSMC		
SMIC	Sony	Toshiba	UMC		
Sony	ST Microelectronics	TSMC			
ST Microelectronics	Texas Instruments	UMC			
Texas Instruments	Toshiba				
Toshiba	TSMC				
TSMC	UMC				
UMC					

Market volume wall: only the largest volume products will be manufactured with the most advanced technology

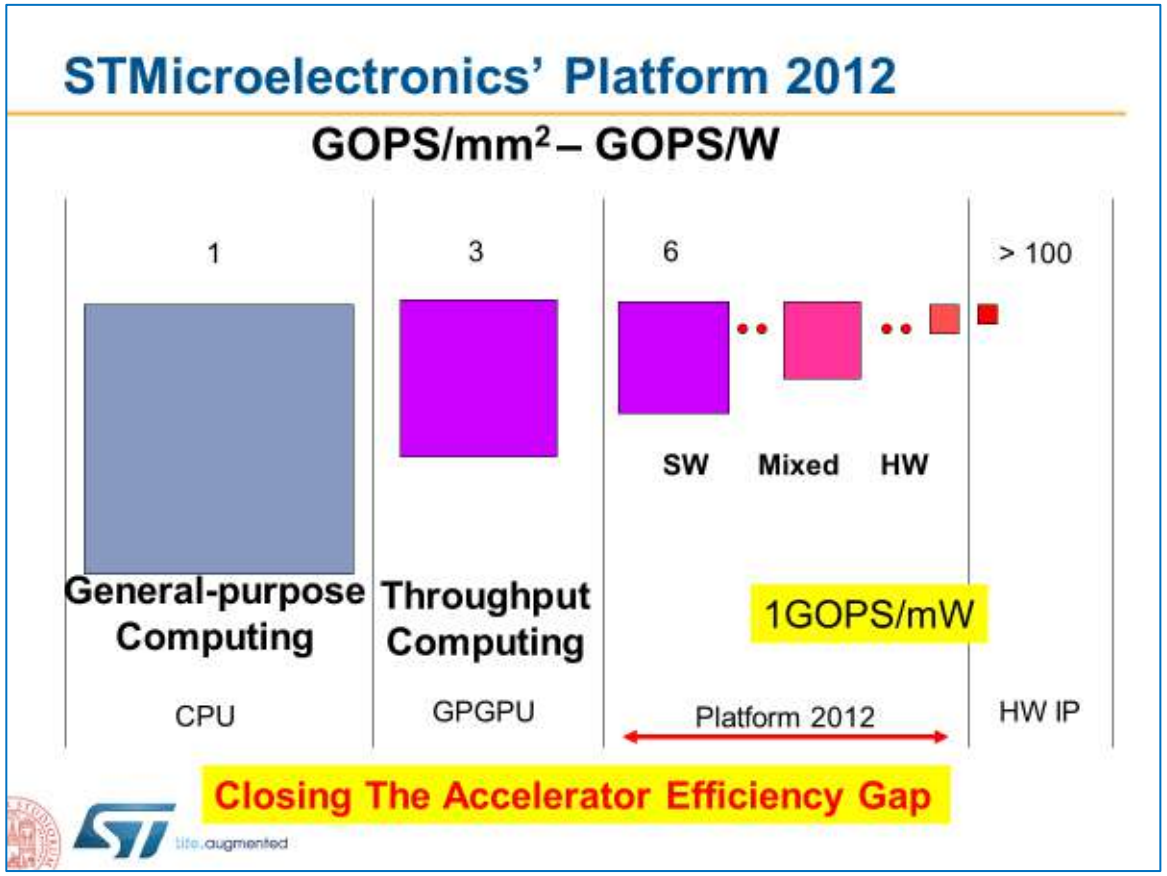
The twilight of Moore's Law: IO Bandwidth

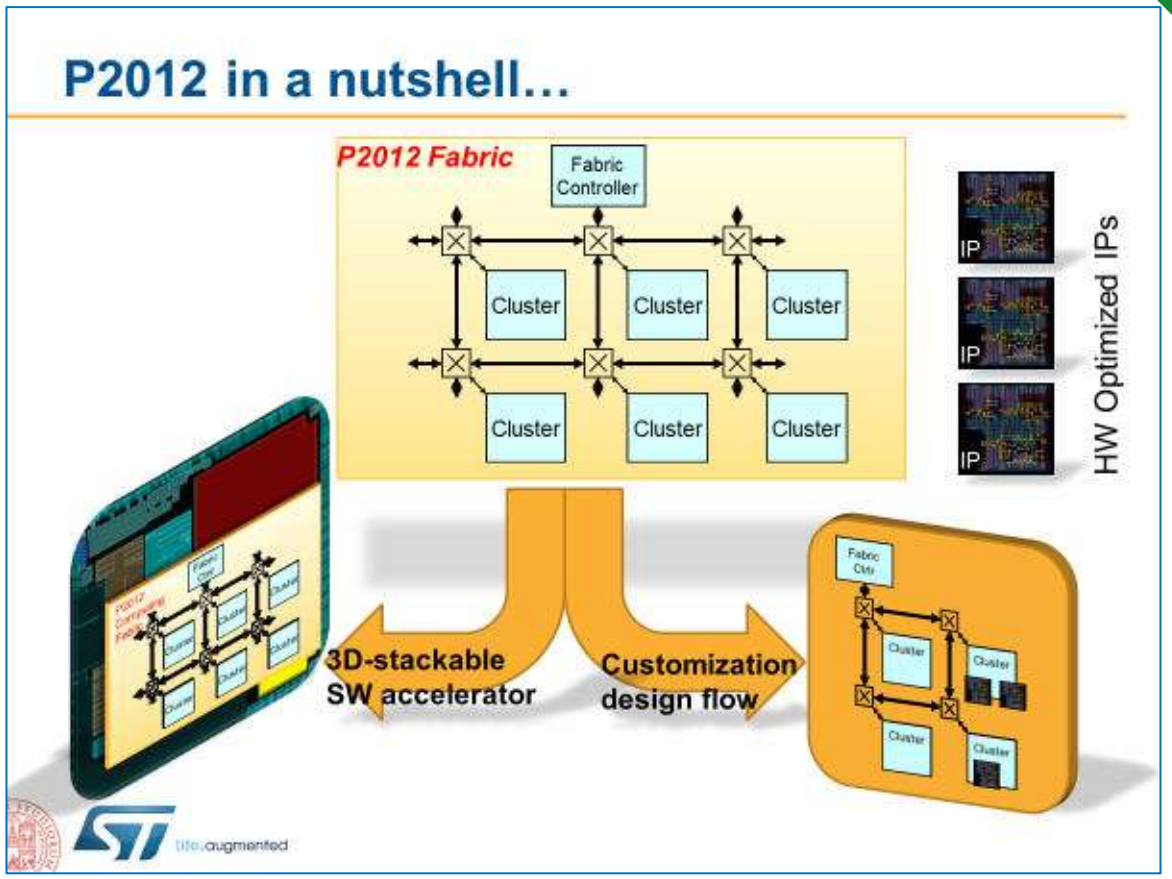


Memory wall: larger datasets and limited bandwidth at high power cost for accessing external memory

ETH zürich ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

Looking Back: April 2012, Job Talk @ ETHZ



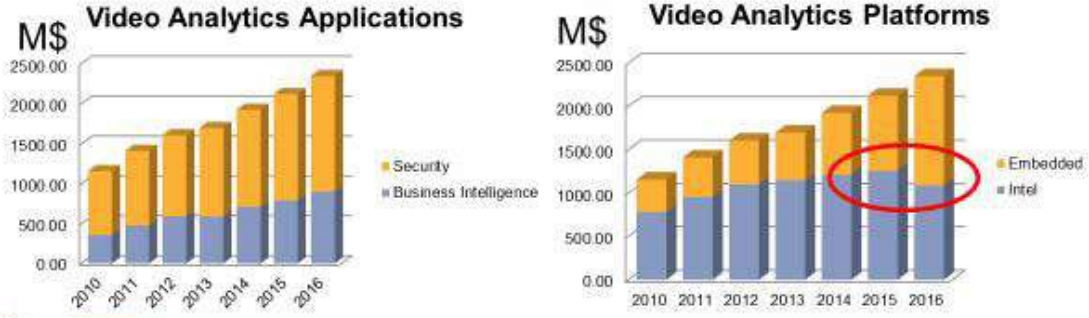


Heterogeneous, Accelerated Computing, 3D integration…

Looking Back: April 2012, Job Talk @ ETHZ

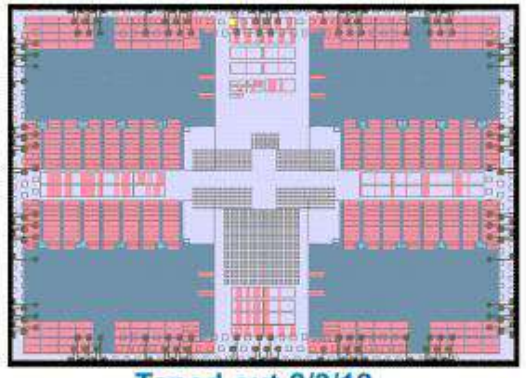
A Killer Application (domain) for P2012

Embedded Visual intelligence



The next killer app: Machines that see (J. Bier)

P2012 SoC in 28nm



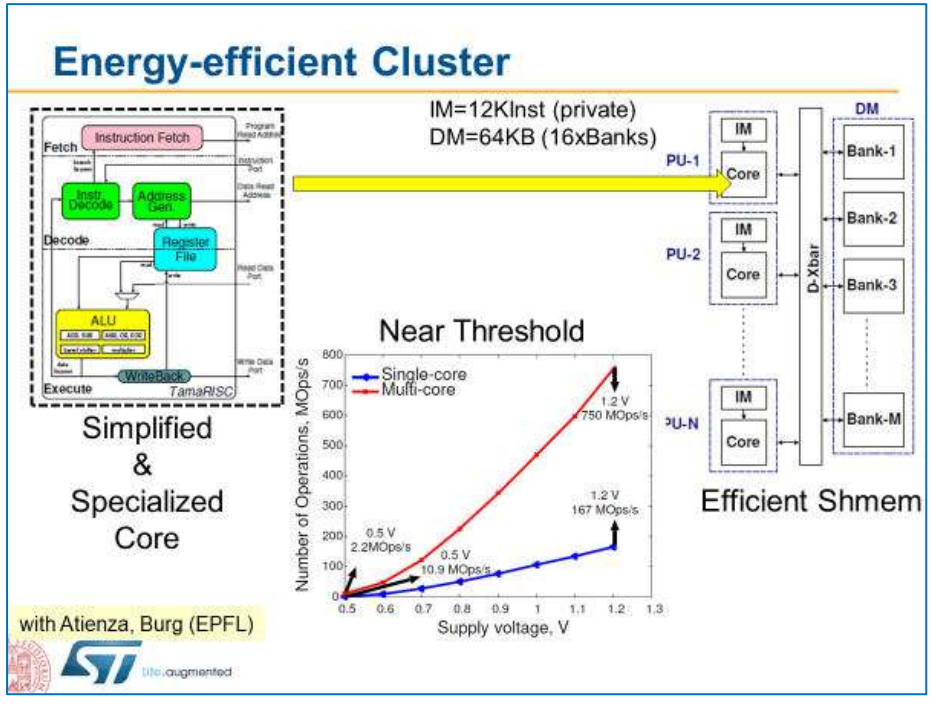
Taped out 2/3/12

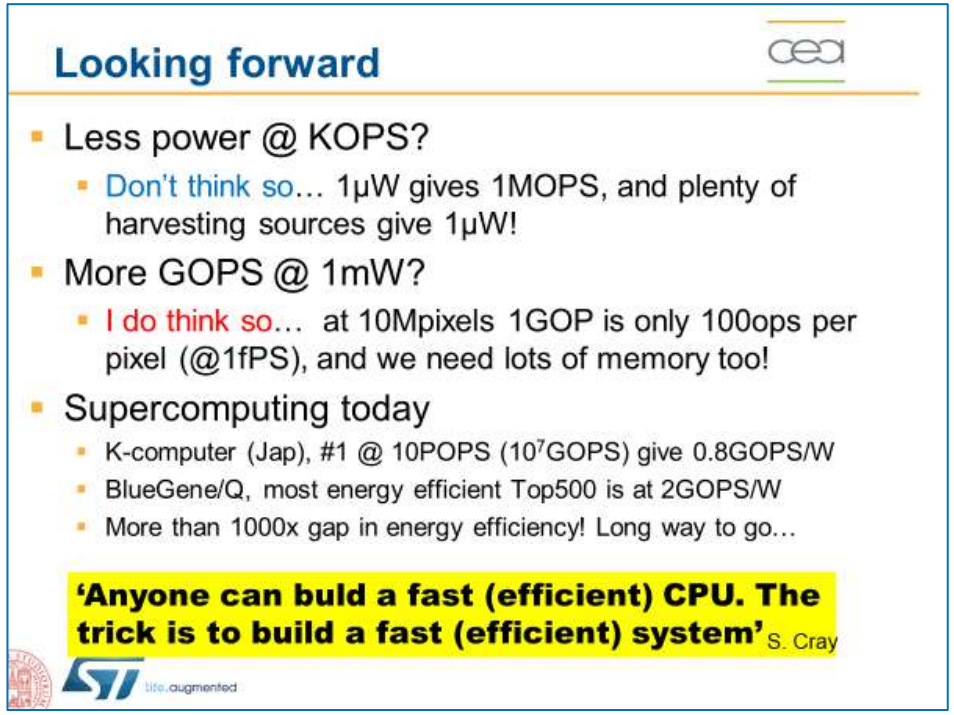
- 4 Clusters, 69 processors
- 80 GFlops
- 1MB L2 mem
- 2D flip chip or 3D stacked
- 600 MHz typ
- < 2 W
- 3.7 mm2 per cluster

Energy efficiency 40GOPS/W →0,04GOPS/mW

P2022 was too early + Crashed against ARM dominance

Looking Back: A few good ideas





Parallel, Ultra Low Power Processors

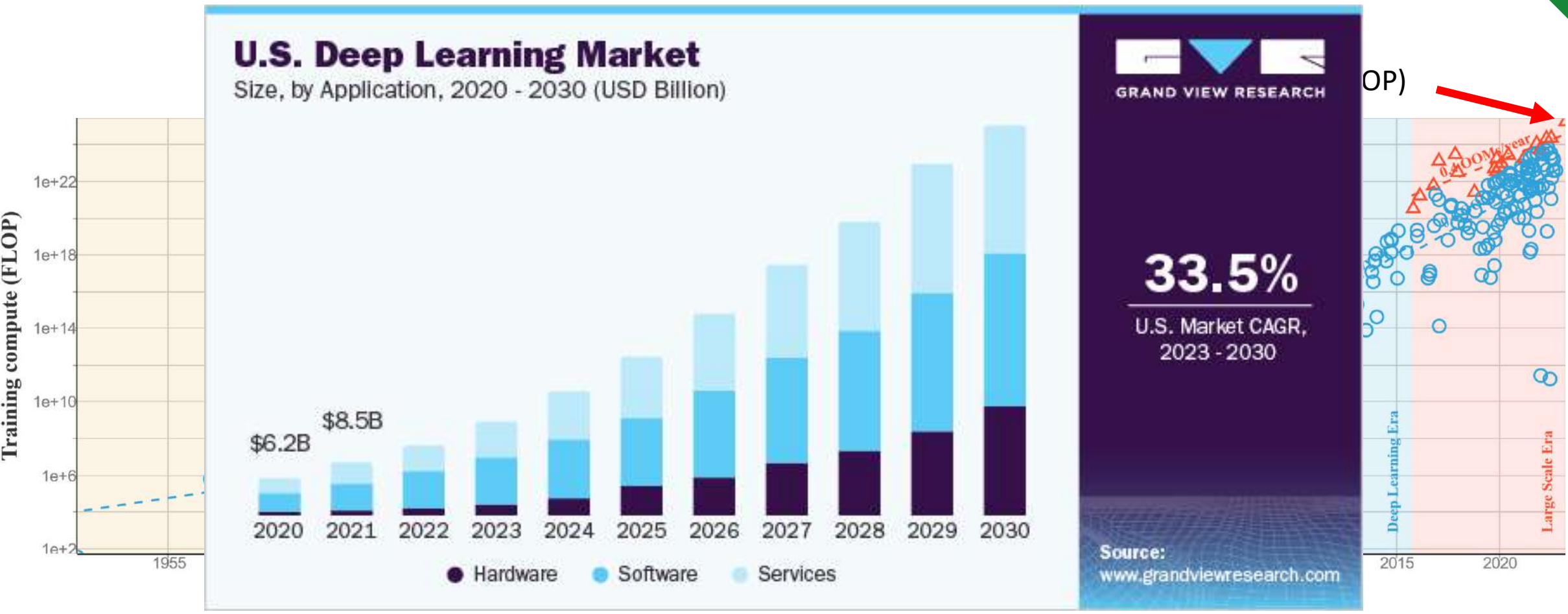
Target pJ/OP @ GOPS and beyond

But for what?

And how to escape the proprietary ISA cage?

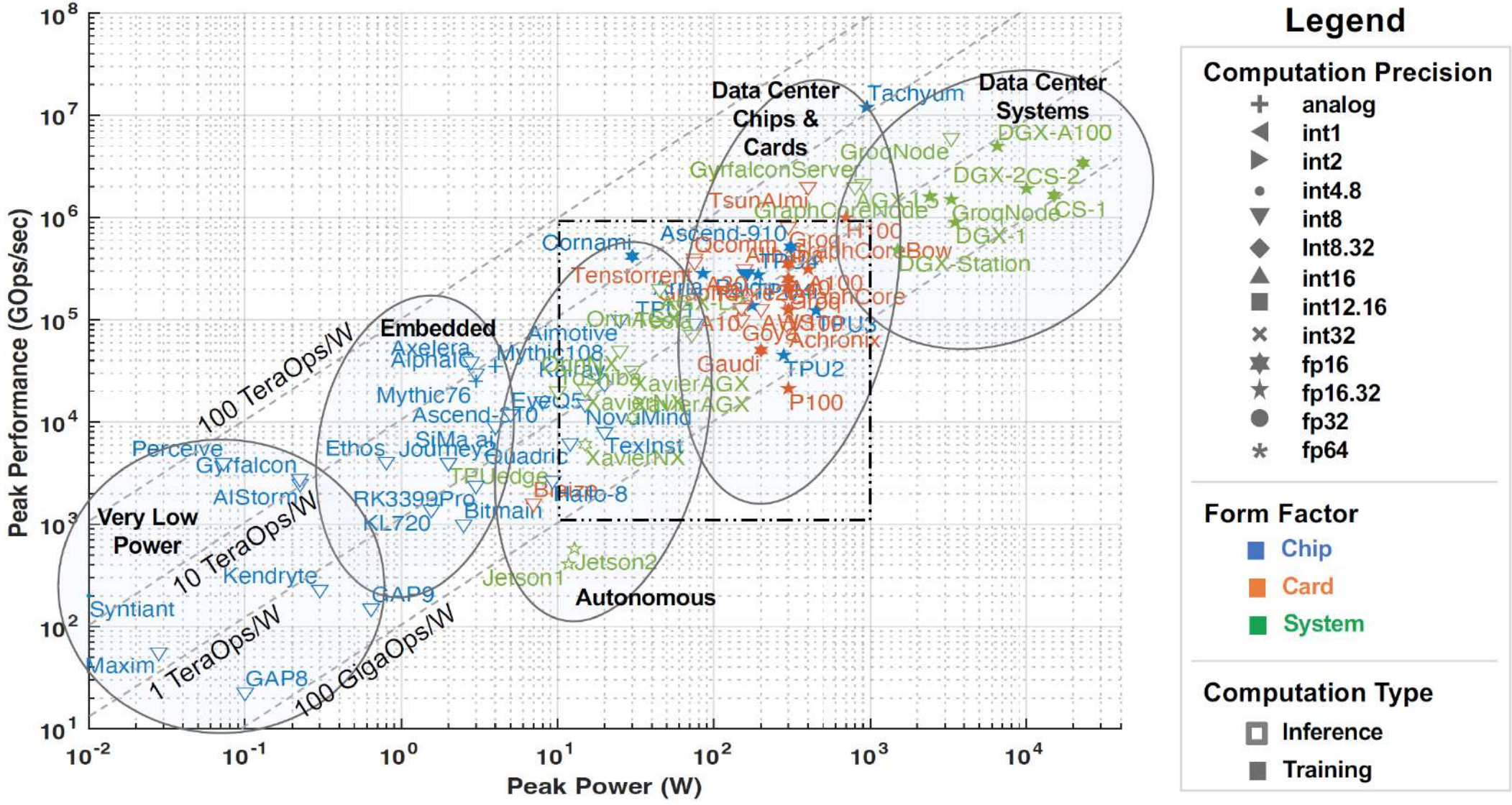
Looking Back: Serendipity!

Sevilla 22: arXiv:2202.05924, epochai.org



10x every 2 years

Looking Back: Serendipity!



Reuther 22: arXiv:2210.04055

Looking Back: More Serendipity!

2014 – A cute, open ISA

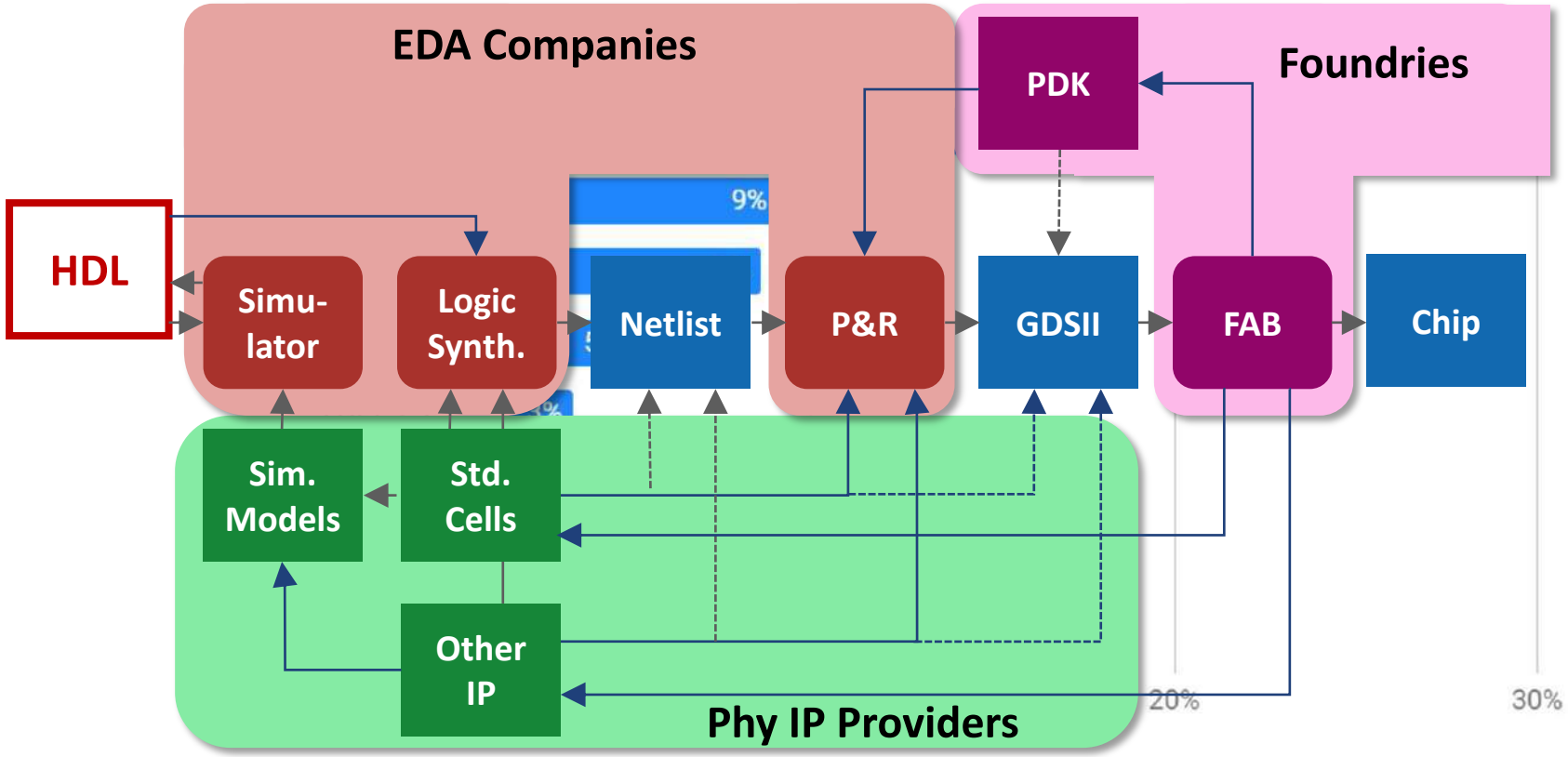
2023 – Disruptive Force

2023 RISC-V International more than 26% membership growth year-over-year, with over 3,180 members across 70 countries. More than 10 billion RISC-V cores in the market, 10K+ engineers working on RISC-V

A Good Intuition

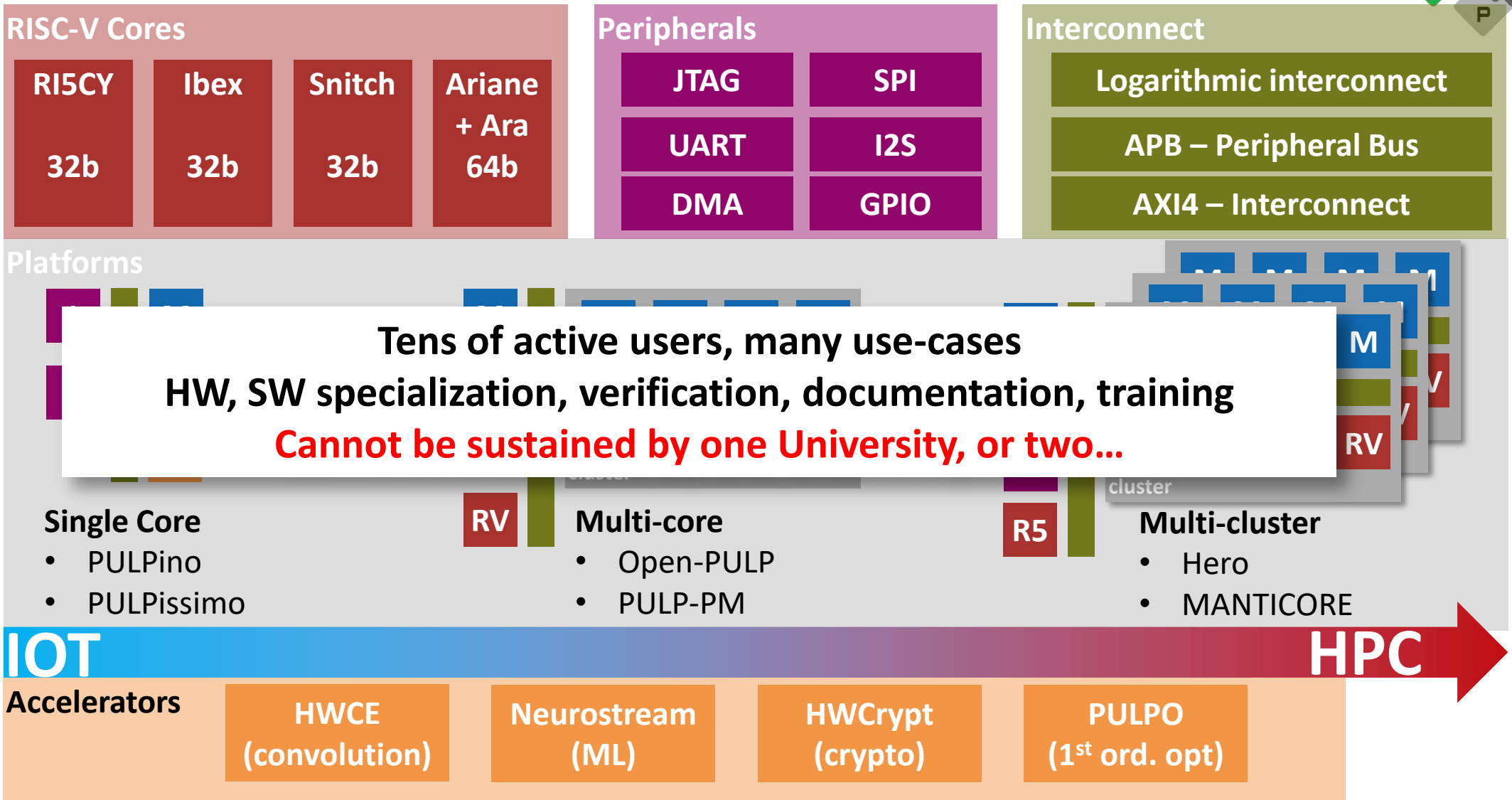
Open Source Hardware! → RTL source code (permissive*, e.g. Apache is key for industrial adoption)

Later stages contain closed IP of various actors → not open source by default (working on that…)



"See: https://cern-ohl.web.cern.ch/ (CERN-OHL-S, -W, -P)

Open Source Platform



A Fast Growing Industrial Open Source Ecosystem

OPENHW GROUP™ PROVEN PROCESSOR IP

Rick O'Connor (OpenHW CEO, former RISC-V foundation director)

- **OpenHW Group** is a not-for-profit, global organization (EU,NA,Asia) where HW and SW designers collaborate in the development of open-source cores, related IP, tools and SW such as the **Core-V** family
- **OpenHW Group** provides an infrastructure for hosting high quality open-source HW developments in line with industry best practices.

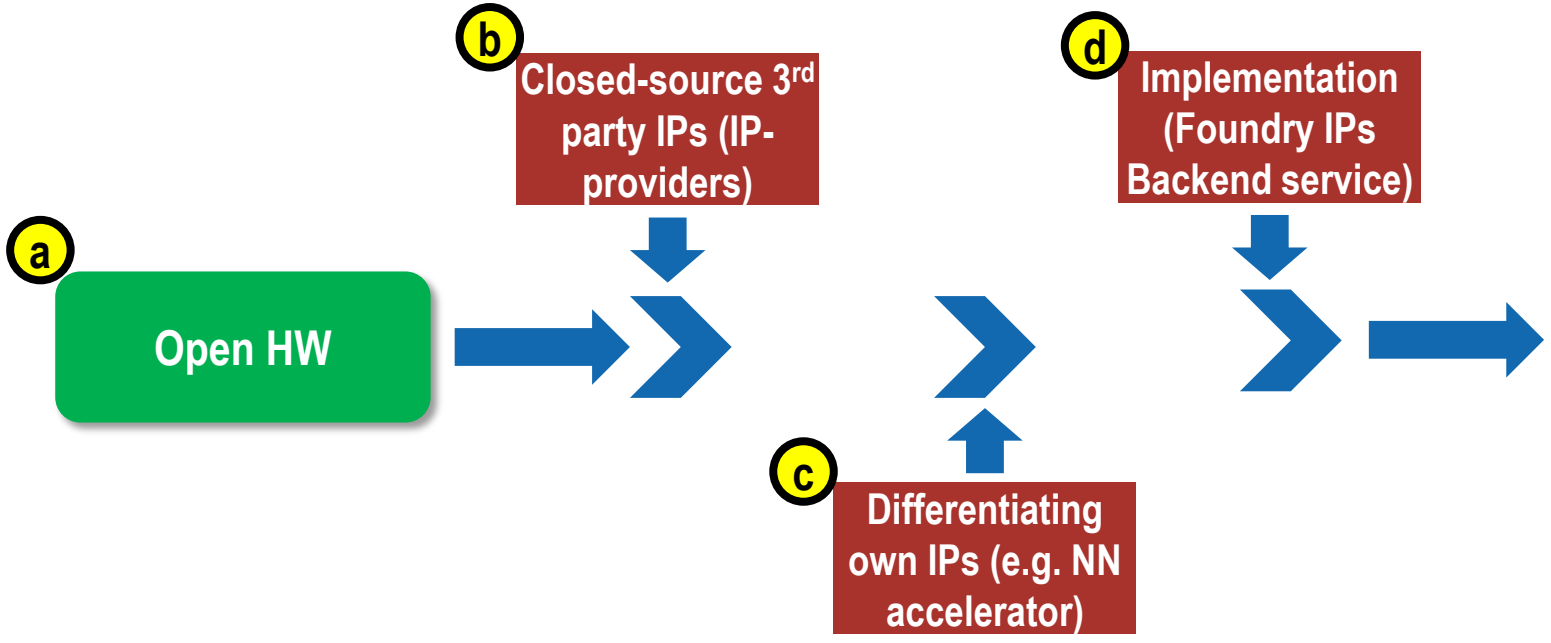
PULP Parallel Ultra Low Power

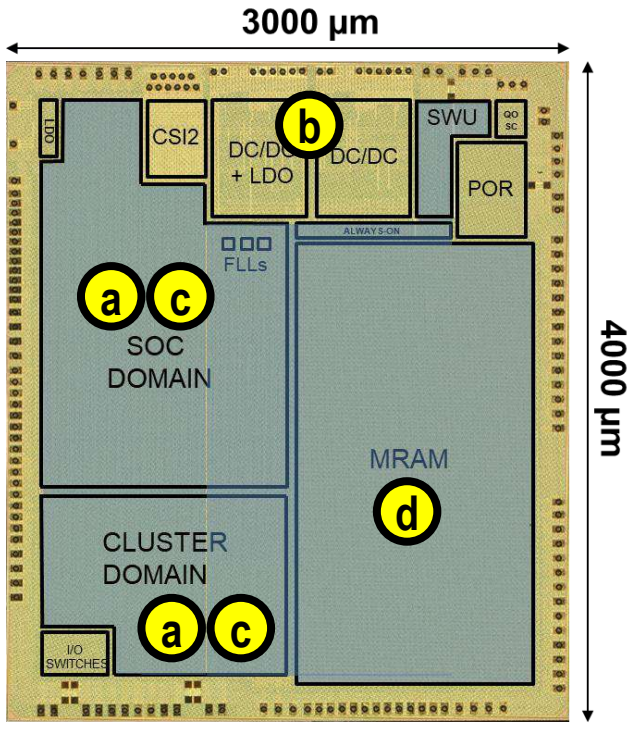
RI5CY, ARIANE

CORE-V™

80+ members today

Creating Product Value with OSHW

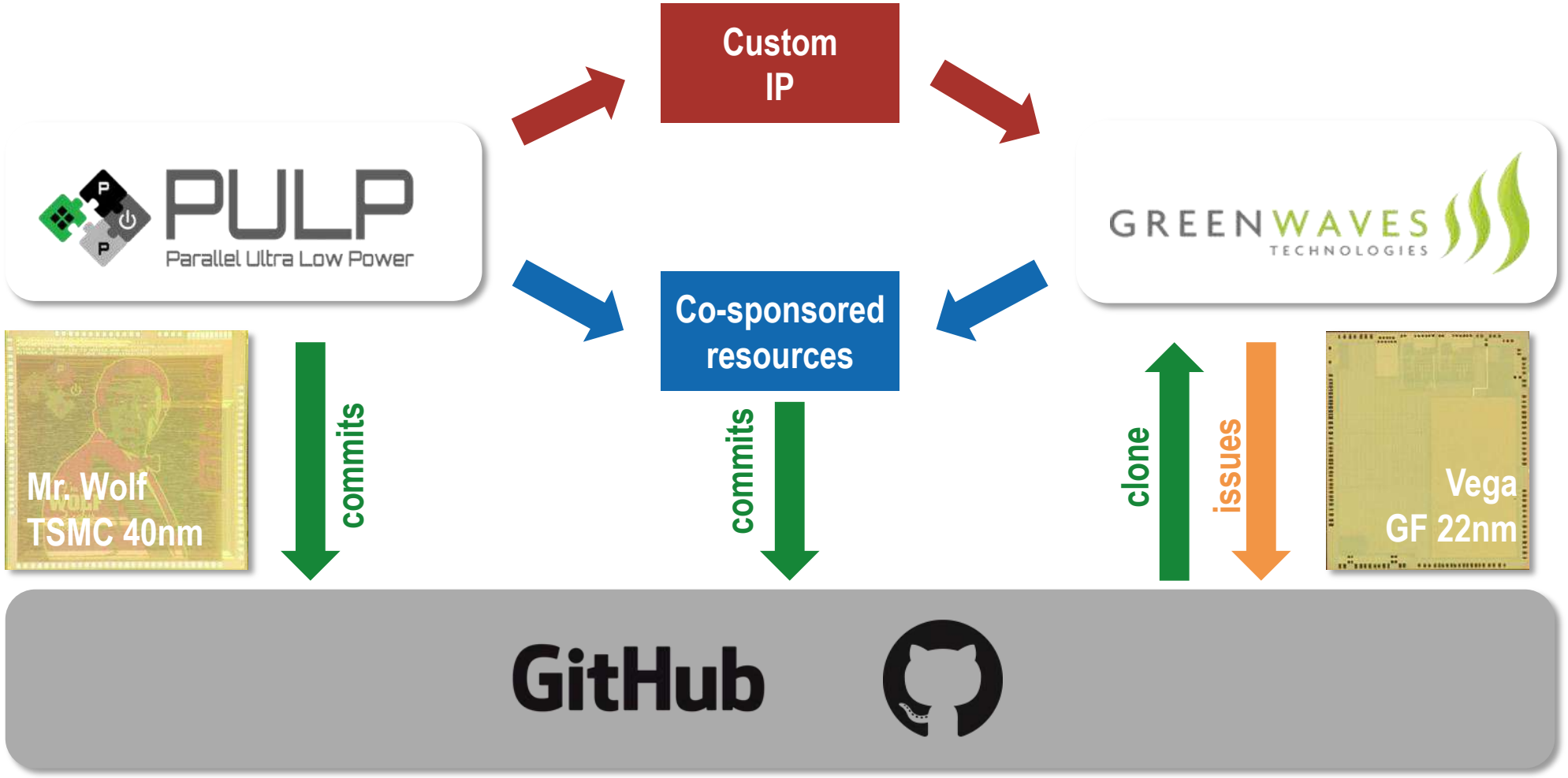




First iteration: test-chip for IP qualification, early customer engagement (MPW)
Second iteration: first low volume production (most effort on c and d) (MLR or full mask set)

NOTE – **aggressive** (e.g. Greenwaves: IoT processor) **vs.** **cost-sensitive** fabless (e.g Eggtronics: cellular charger IC) users
Aggressive: customizing OSHW to provide differentiation wrt to ARM (differentiation). Targets advanced nodes
Cost-sensitive: using OSHW "as is" to reduce cost wrt to ARM, and TtM, effort wrt to in-house, Targets older nodes

With a Little Help from my Friends…



Now to Eric+Loic!

Forward to 2022: Job Done?

Forward to 2022: Job Done?

Forward to 2022: Job Done?

RISC-V Cluster

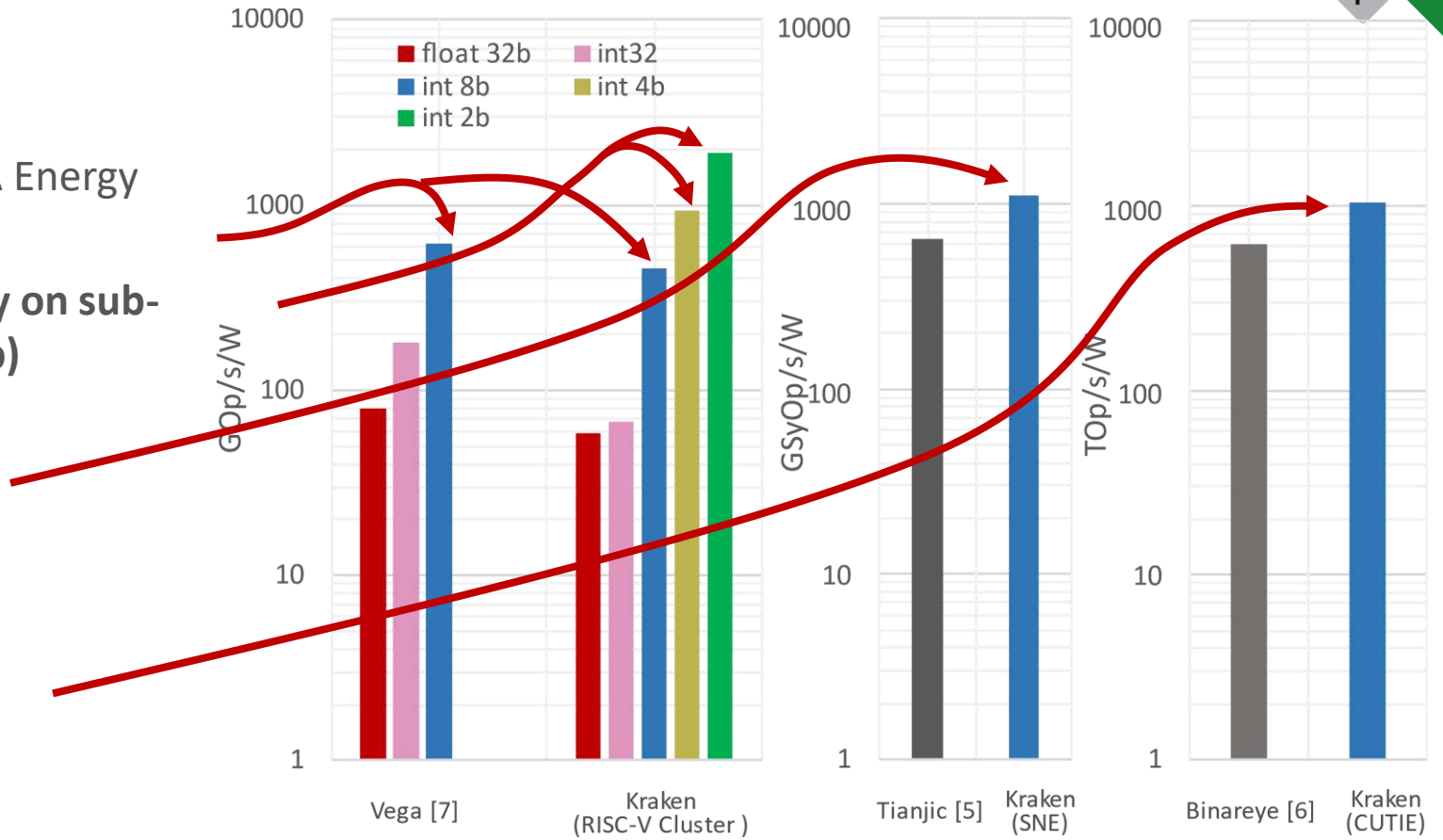
- Comparable 32bits-8bits SOA Energy efficiency to other PULPs [7]
- **The highest energy efficiency on sub-byte SIMD operations (4b-2b)**

SNE

- **1.7X** higher than SOA [5] energy/efficiency

CUTIE

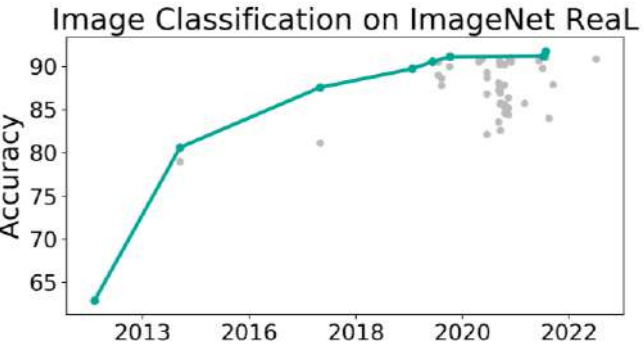
- **2X** higher energy efficiency improvement over SOA [6]

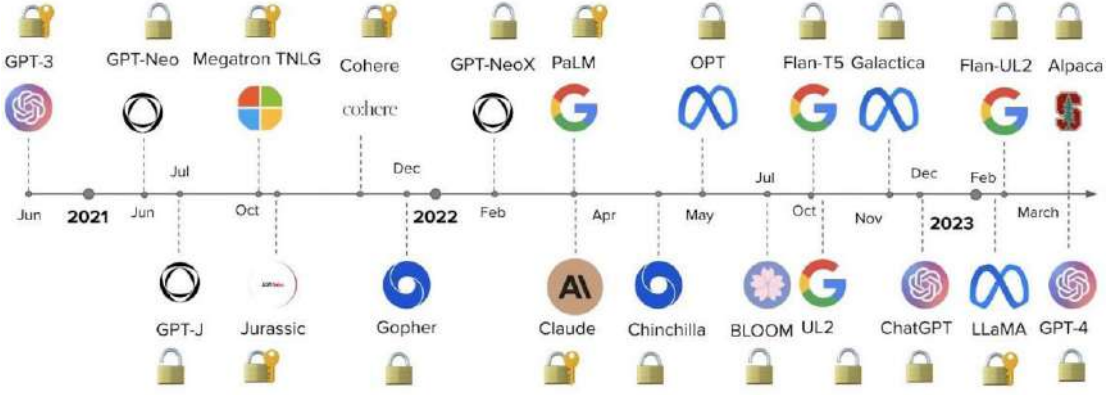


CUTIE, SNE can work concurrently for SNN + TNN "fused" inference (never done so far)

[5] L. Deng et al., "Tianjic," JSSC 2020
[6] B. Moons et al., "Binareye," CICC, 2018
[7] D. Rossi et al., "Vega," JSSC 2022.

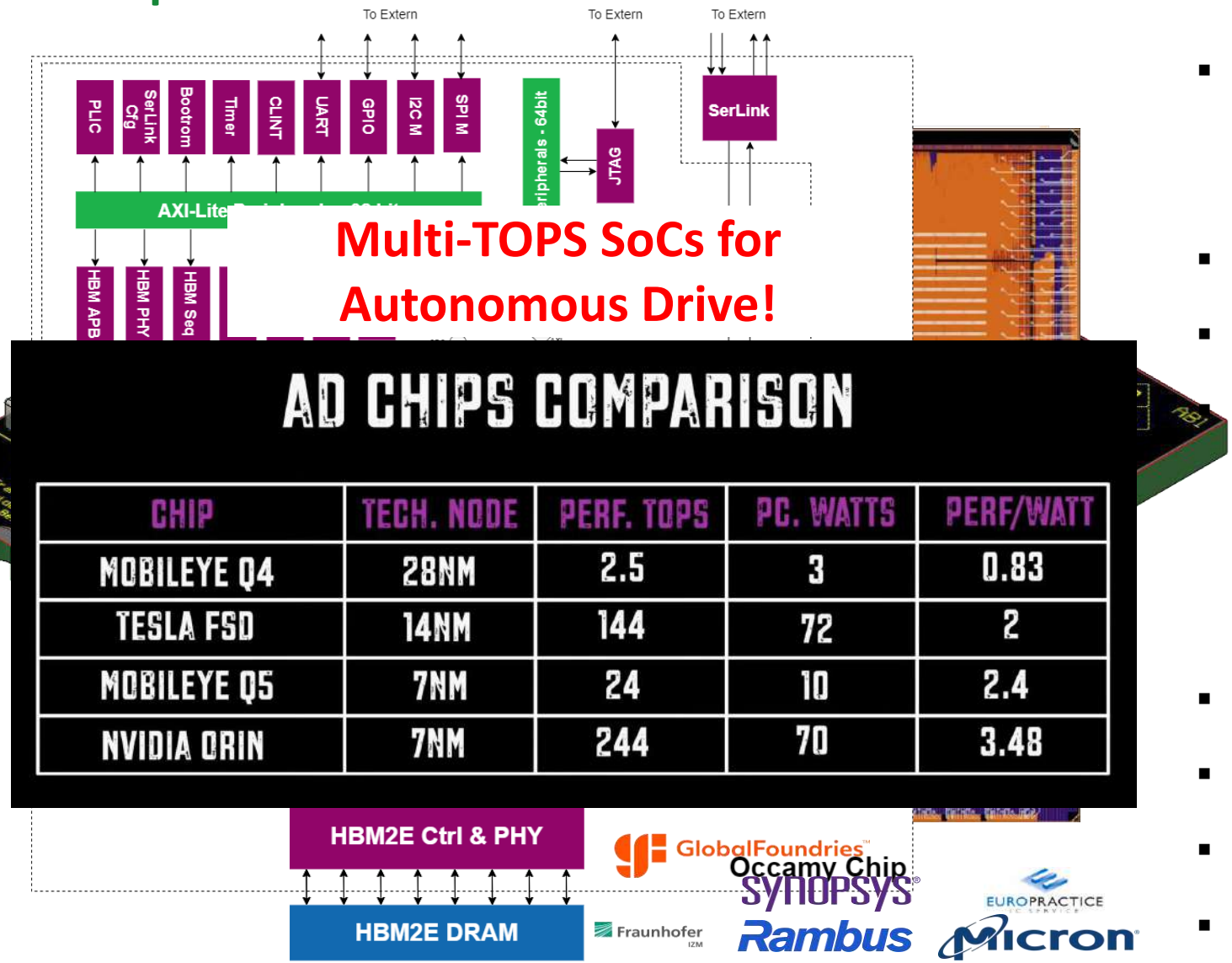
Fast Forward: Perceptive → Generative → Embodied AI





Precise **Interactive, creative** **Efficient, RT-safe, secure**

Disruptive Embodied AI: Automotive



AD CHIPS COMPARISON

CHIP	TECH. NODE	PERF. TOPS	PC. WATTS	PERF/WATT
MOBILEYE Q4	28NM	2.5	3	0.83
TESLA FSD	14NM	144	72	2
MOBILEYE Q5	7NM	24	10	2.4
NVIDIA ORIN	7NM	244	70	3.48

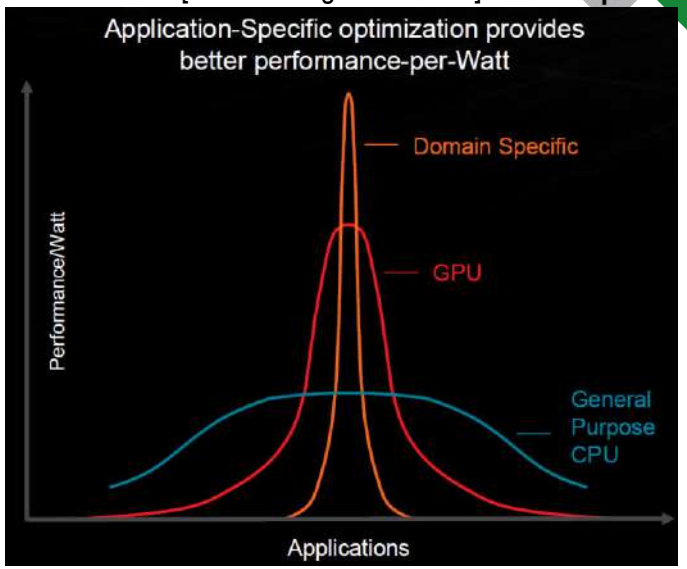
- **GF12, target 1GHz (typ)**
- **2 AXI NoCs (multi-hierarchy)**
  - 64-bit
  - 512-bit with “interleaved” mode
- **Peripherals**
- **Linux-capable manager core CVA6**
- **6 Quadrants: 216 cores/chiplet**
  - 4 cluster / quadrant:
    - 8 compute +1 DMA core / cluster
    - 1 multi-format FPU / core (FP64,x2 32, x4 16/alt, x8 8/alt)
- **8-channel HBM2e (8GB) 512GB/s**
- **D2D link (Wide, Narrow) 70+2GB/s**
- **System-level DMA**
- **SPM (2MB wide, 512KB narrow)**

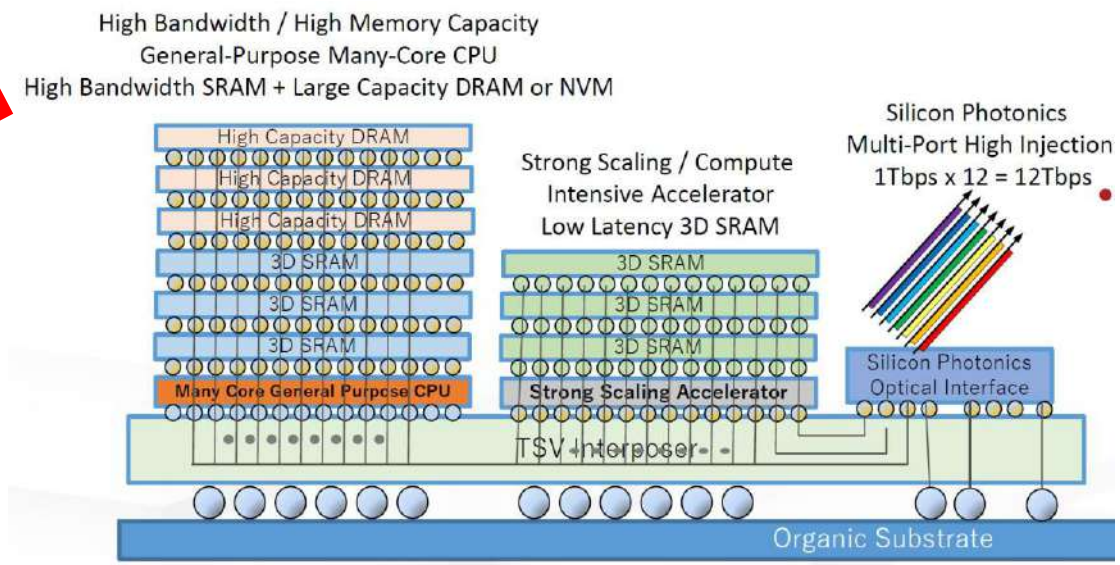
Peak 384 GDPflop/s per chiplet

Conclusion

- Efficient, RT, Safe Secure: PE, Cluster, SoC, System
- Key ideas
  - Deep PE optimization → extensible ISAs (RISC-V!)
  - Low-overhead work distribution. Latency hiding → large “mempools”
  - Heterogeneous architecture →host+accelerator(s)
- Game-changing technologies
  - “Commoditized” chiplets: 2.5D, 3D
  - Computing “at” memory (DRAM mempool)
  - Coming: optical IO and smart NICs, swiches
- Challenges:
  - High performance RV Host
  - RV HPC software ecosystem?
  - Access to technology!

[AMD Naffziger ISCAS22]





[RIKEN Matsuoka MODSIM22]

PULP

Parallel Ultra Low Power

Luca Benini, Alessandro Capotondi, Alessandro Ottaviano, Alessio Burrello, Alfio Di Mauro, Andrea Borghesi, Andrea Cossettini, Andreas Kurth, Angelo Garofalo, Antonio Pullini, Arpan Prasad, Bjoern Forsberg, Corrado Bonfanti, Cristian Cioflan, Daniele Palossi, Davide Rossi, Fabio Montagna, Florian Glaser, Florian Zaruba, Francesco Conti, Georg Rutishauser, Germain Haugou, Gianna Paulin, Giuseppe Tagliavini, Hanna Müller, Luca Bertaccini, Luca Valente, Manuel Eggimann, Manuele Rusci, Marco Guermandi, Matheus Cavalcante, Matteo Perotti, Matteo Spallanzani, Michael Rogenmoser, Moritz Scherer, Moritz Schneider, Nazareno Bruschi, Nils Wistoff, Pasquale Davide Schiavone, Paul Scheffler, Philipp Mayer, Robert Balas, Samuel Riedel, Segio Mazzola, Sergei Vostrikov, Simone Benatti, Stefan Mach, Thomas Benz, Thorir Ingolfsson, Tim Fischer, Victor Javier Kartsch Morinigo, Vlad Niculescu, Xiaying Wang, Yichao Zhang, Frank K. Gürkaynak,
all our past collaborators and many more that we forgot to mention

http://pulp-platform.org @pulp_platform