Head in the Clouds –
Building a chip for scale out computing

Bryan Chin & the Cavium Team

Cavium, Inc.

Agenda

- *Cavium Background*
- *Network Processor*
- *Core characteristics*
- *Data center processor*
- *Potential for performance*

Cavium SoC's for Range of Target Markets

Networking			Consumer	Wireless		Storage	Server
Enterprise & Edge Routers	Enterprise, Metro Switches & L4-L7 Equipment	Security & DPI Equipment	Home, Video, High Bandwidth Broadband	3G, 4G Infra-structure	LAN: Controllers & Enterprise AP's	Storage Networking, Arrays & Adapters	Cloud & Data Centers

Networking : Router, Appliance, Gateway SoCs

OCTEON

Security, Compression, DPI Offload & Virtualization SoCs

NITROX

Video SoCs

PureVu

Media & Set Top SoCs

Celestial

SMB, Home & NAS SoCs

ECONA

Base-station, RAN, Core SoCs

OCTEON

Small Cell Base-station

WLAN Controller & AP SoCs

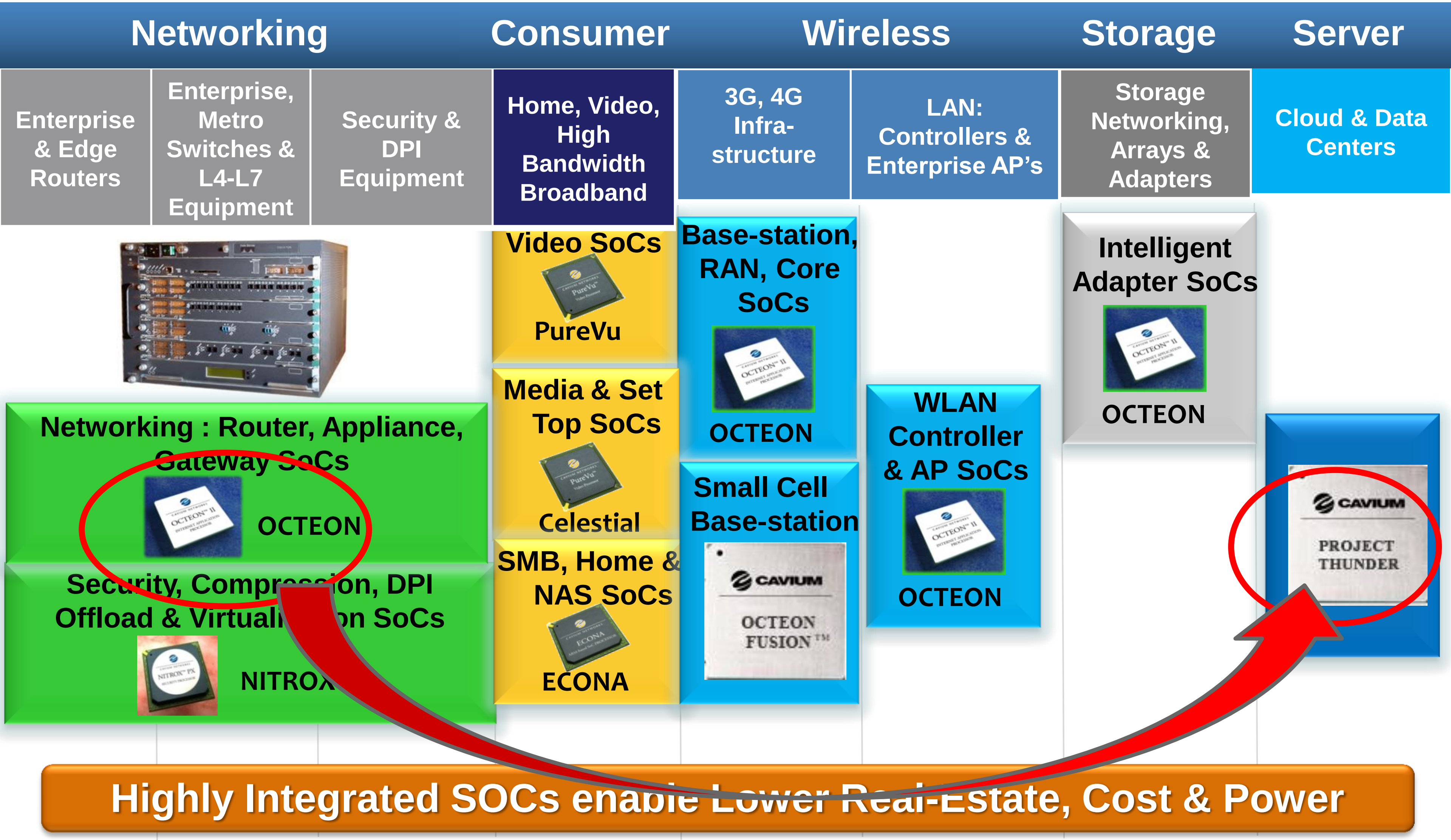
OCTEON

Intelligent Adapter SoCs

OCTEON

Highly Integrated SOCs enable Lower Real-Estate, Cost & Power

The Road from Here to There



Networking | Consumer | Wireless | Storage | Server

Enterprise & Edge Routers

Enterprise, Metro Switches & L4-L7 Equipment

Security & DPI Equipment

Home, Video, High Bandwidth Broadband

3G, 4G Infra-structure

LAN: Controllers & Enterprise AP's

Storage Networking, Arrays & Adapters

Cloud & Data Centers

Networking : Router, Appliance, Gateway SoCs

OCTEON

Security, Compression, DPI Offload & Virtualization SoCs

NITROX

Video SoCs

PureVu

Media & Set Top SoCs

Celestial

SMB, Home & NAS SoCs

ECONA

Base-station, RAN, Core SoCs

OCTEON

Small Cell Base-station

WLAN Controller & AP SoCs

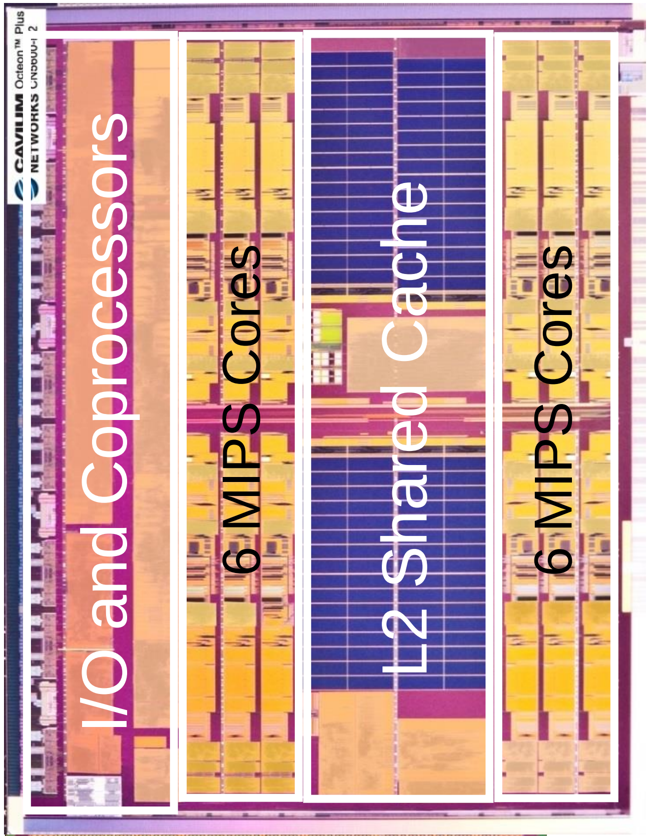
OCTEON

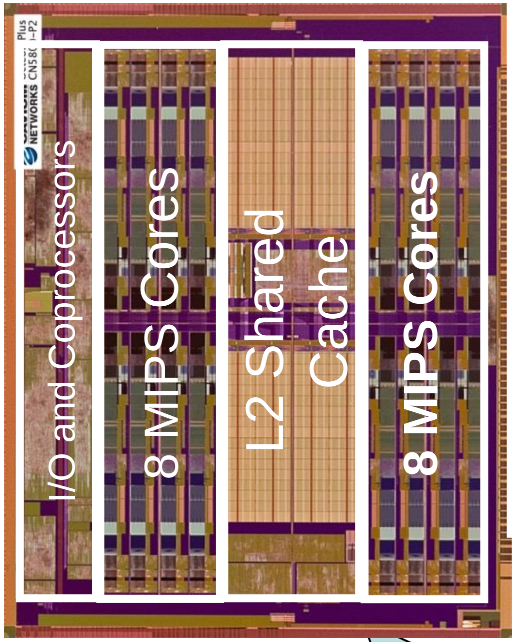
Intelligent Adapter SoCs

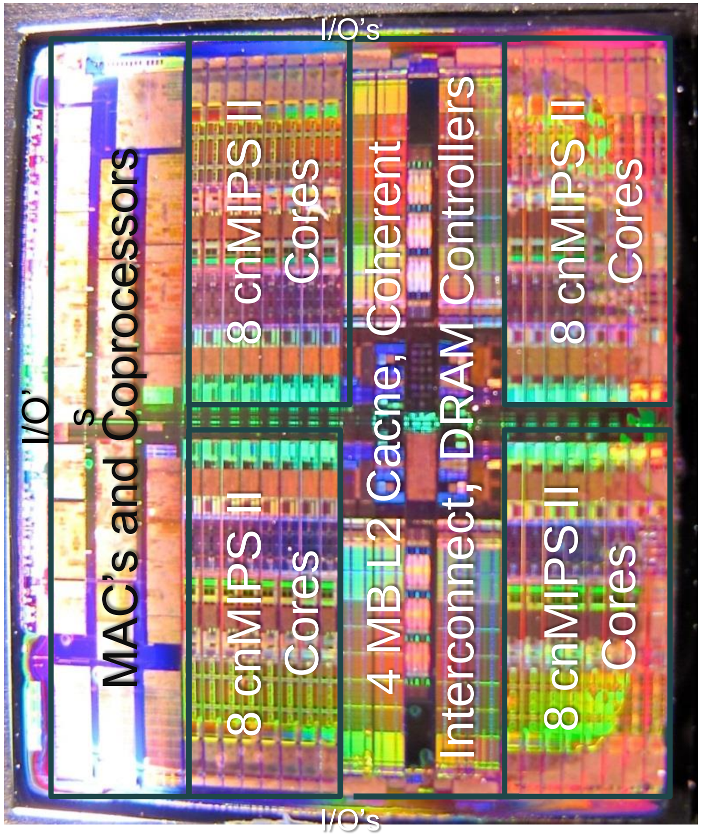
OCTEON

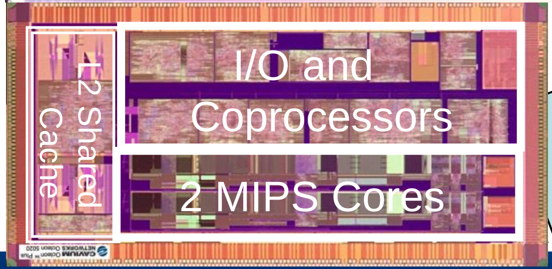
Highly Integrated SOCs enable Lower Real-Estate, Cost & Power

Some Existing OCTEON chips









OCTEON III CN78XX

Cores:

- 48 cnMIPS III @ 2.5GHz
- Large shared L2 Cache w/ ECC
- Cores, Crossbar, L2 @ 2.5GHz

Memory Controllers:

- 4 x 72b DDR3-2133 & DDR4 w/ ECC
- Up to 256GB, 4-rank x4 DIMMs

OCTEON Coherent Interconnect

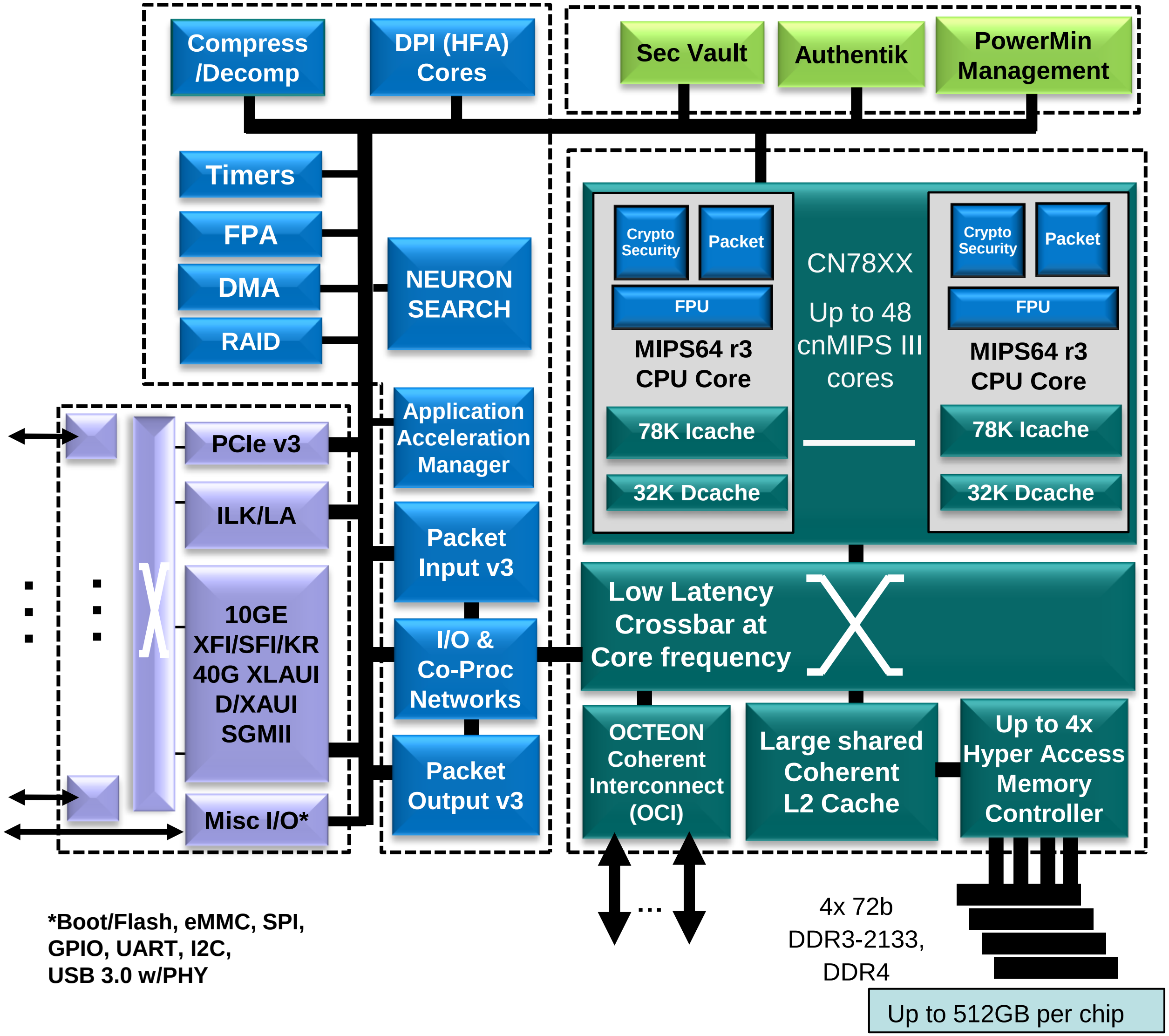
50+ lanes 10+ Gb Serdes

HW Acceleration (Up to 100G+)

- Packet Processing, QoS, TCP, SCTP, MPLS, FCoE, iSCSI
- Packet Ordering, Schedule, Synch.
- Security, Compression
- Deep Packet Inspection (HFA)
- Search and ACL Lookup (NEURON)
- RAID, De-Dup

Compatibility

- Backward and Software compatible with all OCTEON families



*Boot/Flash, eMMC, SPI, GPIO, UART, I2C, USB 3.0 w/PHY

OCTEON III CN78XX

Cores:

- 48 cnMIPS III @ 2.5GHz
- Large shared L2 Cache w/ ECC
- Cores, Crossbar, L2 @ 2.5GHz

Memory Controllers:

- 4 x 72b DDR3-2133 & DDR4 w/ ECC
- Up to 256GB, 4-rank x4 DIMMs

OCTEON Coherent Interconnect

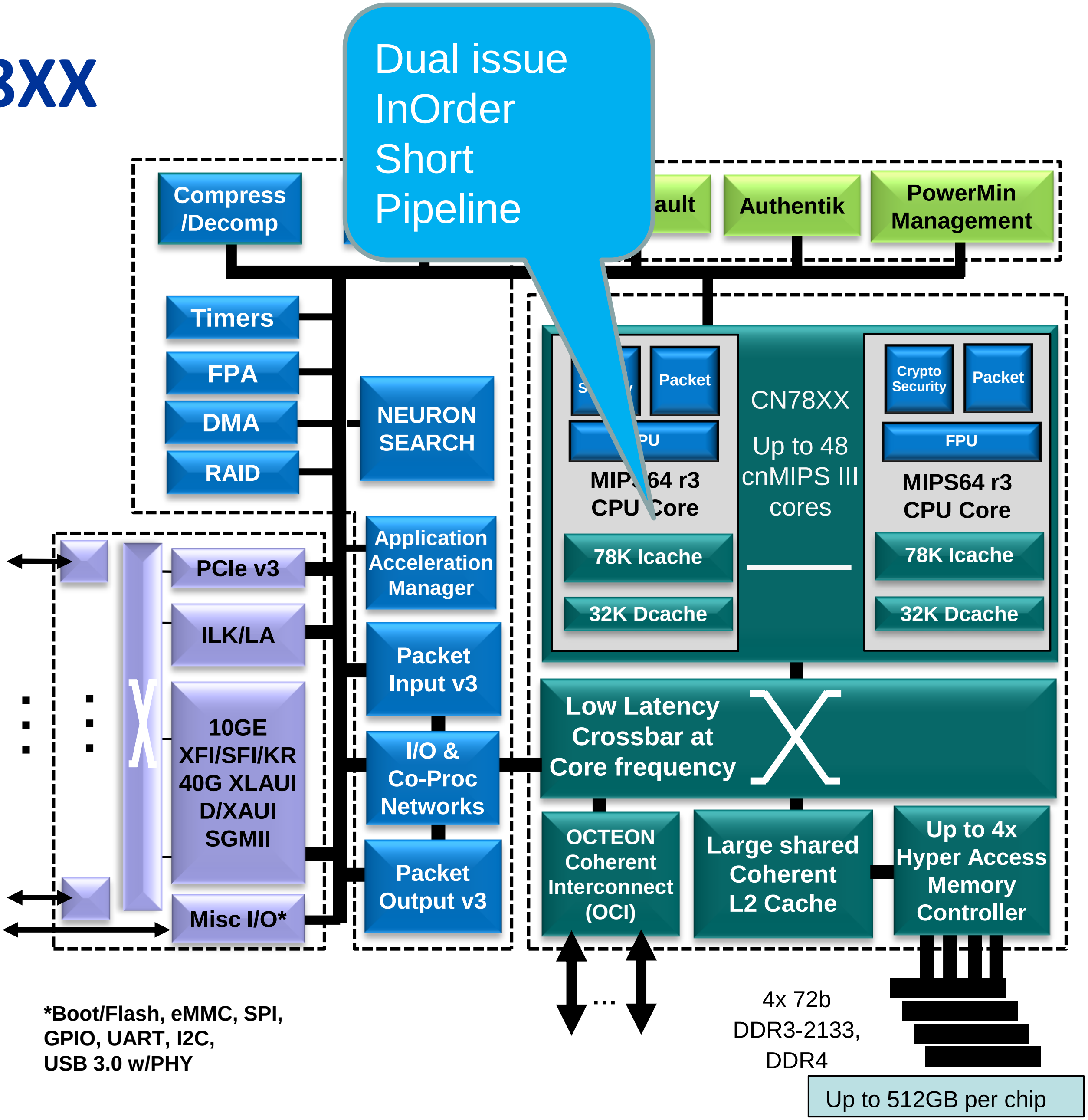
50+ lanes 10+ Gb Serdes

HW Acceleration (Up to 100G+)

- Packet Processing, QoS, TCP, SCTP, MPLS, FCoE, iSCSI
- Packet Ordering, Schedule, Synch.
- Security, Compression
- Deep Packet Inspection (HFA)
- Search and ACL Lookup (NEURON)
- RAID, De-Dup

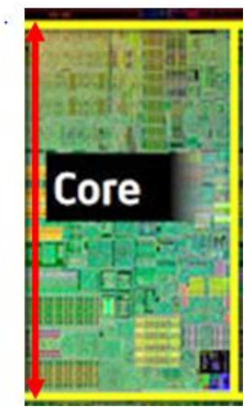
Compatibility

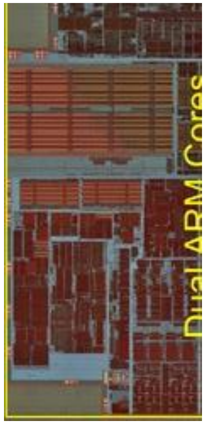
- Backward and Software compatible with all OCTEON families

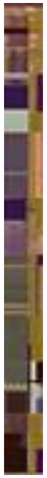


*Boot/Flash, eMMC, SPI, GPIO, UART, I2C, USB 3.0 w/PHY

Relative Sizes







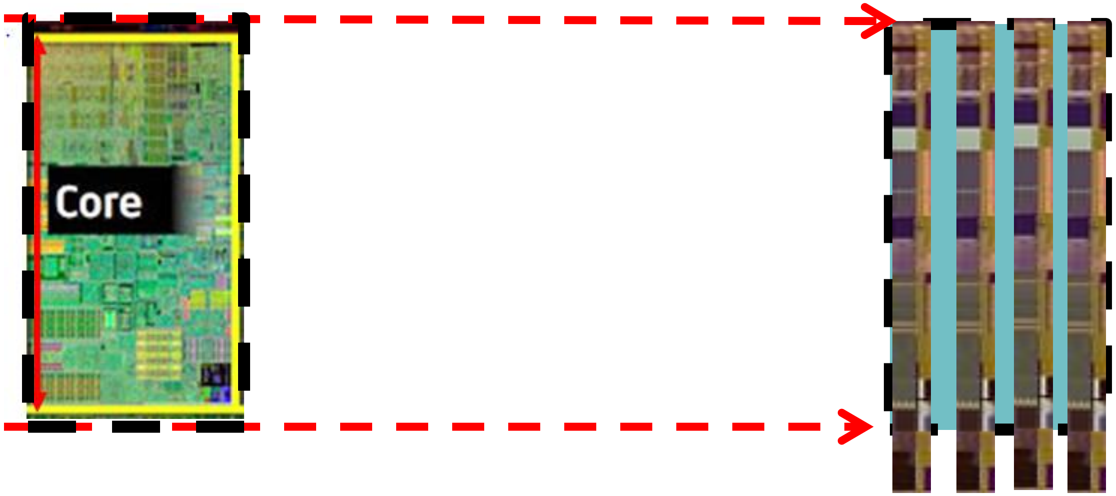
Sandybridge
4 decode
6 issue
OoO

Apple A7
6 issue
OoO

Cavium MIPS core
2 issue
In order

It's hard to get interesting die photos because of all the metal…

Angels on a Pinhead



- Can fit about 4 simple cores in area of one complex core
- But what is the cost?
  - Simpler micro-architecture (less ILP, MLP)
  - Shallow memory hierarchy
  - Global shared L2 instead of private or semiprivate L2

OCTEON II L1d (DCACHE) MPKI Generally

Less Than Nehalem L2 (MLC) MPKI

Nehalem – 32K 4 way I, 32KB 8 way D, 256K L2, 4-24 MB
OCTEON II – 37KB, 37 way I, 32KB 32 way D, 4MB 16 way L2

MPKI = Misses Per Kilo Instructions
MLC = Mid Level Cache

SPEC2006 Integer	32KB L1d MPKI Cavium Octeon II	256KB L2 MPKI Intel Nehalem
401.bzip2	6.21	8.34
429.mcf	84.59	108.1
445.gobmk	2.11	3.03
456.hmmer	1.08	3.02
458.sjeng	1.14	0.89
462.libquantum	12.91	38.6
464.h264ref	3.99	2.25
473.astar	9.49	10.6

Highly associative 1st level cache same or better hit rate than 2 level Private Cache

Characteristics of the Workloads

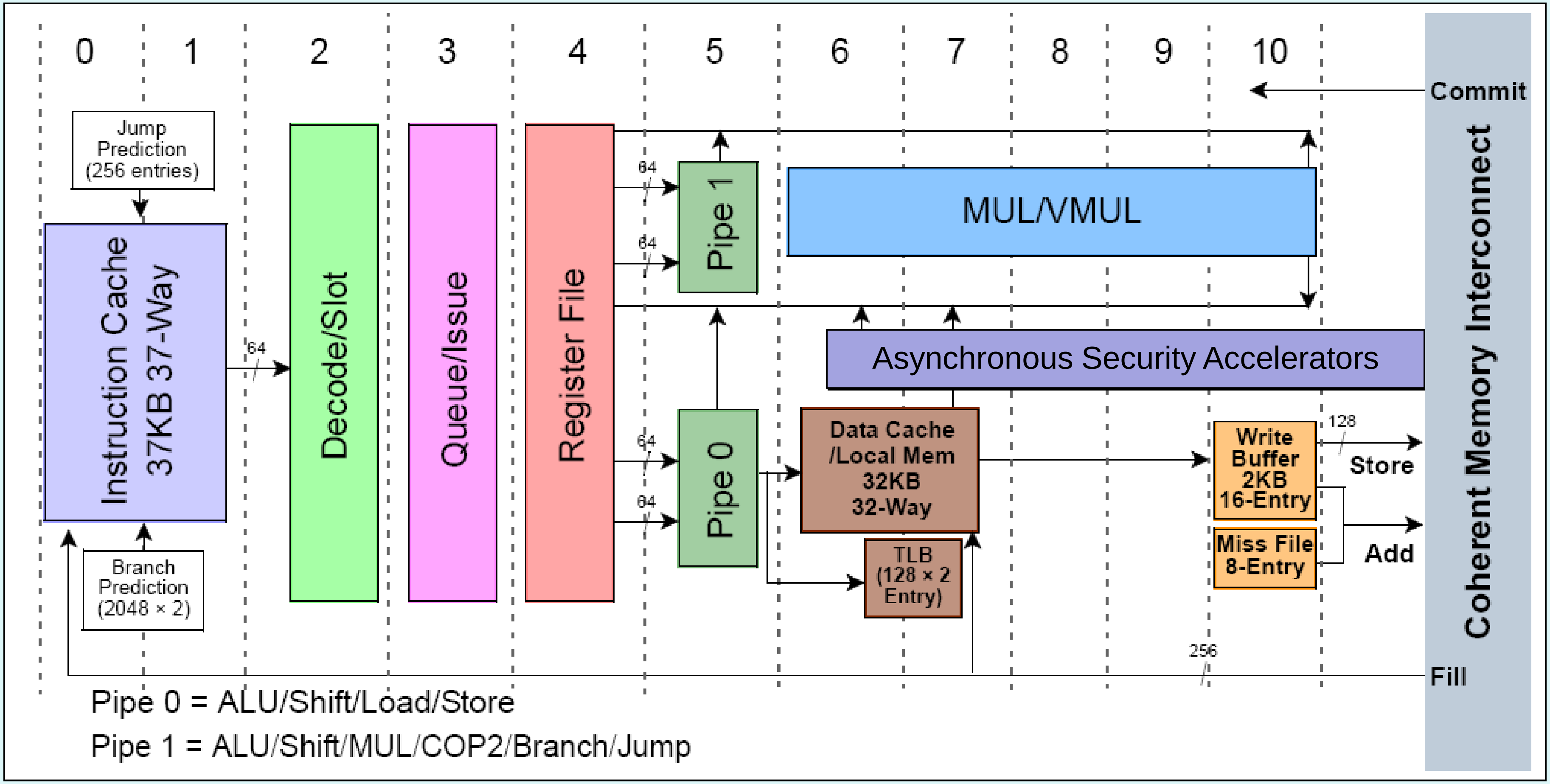
	Networking	Scaleout	Enterprise
Highly parallel	✓	✓	
Benefit from ILP/MLP (OoO) *			✓
Repetitive task on lots of data	✓	✓	
Hardware accelerator Friendly	✓	✓	Sometimes (e.g. GPU)
Compile once, run many	✓	✓	

* Ferdman et al; “Clearing the Clouds: A Study of Emerging Scale-Out Workloads on Modern Hardware; ASPLOS 2012

In this case, simpler is “better”

- Cloud workloads -> Not a lot of ILP or MLP
- *What can you do to improve performance?*
- OoO depends on
  - – Available ILP
  - – Overlapping independent memory operations (MLP)
- Build an in order machine
  - – Reduce the load to use latency in the L1-Dcache

cnMIPS II Core 8+ Stage Pipeline

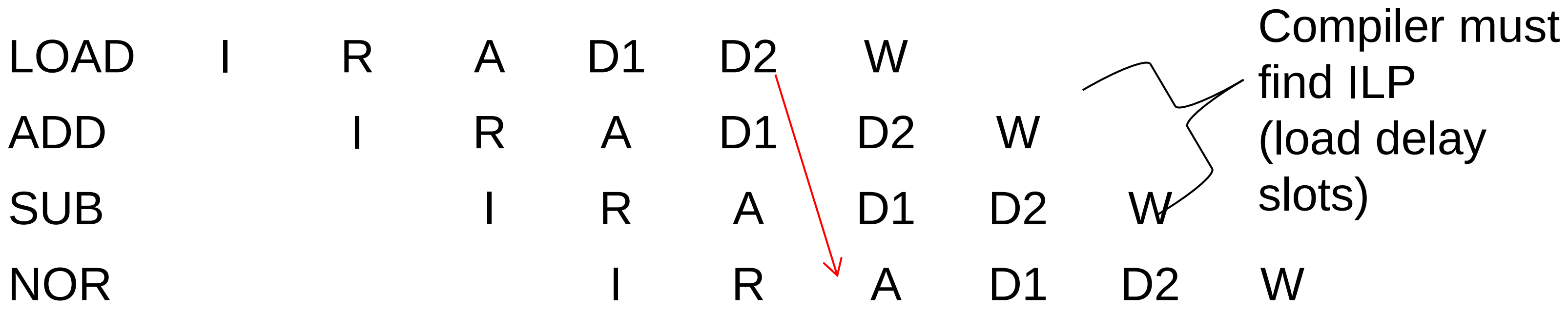

0
1
2
3
4
5
6
7
8
9
10
Commit
Jump Prediction (256 entries)
Instruction Cache 37KB 37-Way
Branch Prediction (2048 × 2)
64
Decode/Slot
Queue/Issue
Register File
64
64
Pipe 1
64
64
Pipe 0
MUL/VMUL
Asynchronous Security Accelerators
Data Cache /Local Mem 32KB 32-Way
TLB (128 × 2 Entry)
Write Buffer 2KB 16-Entry
Miss File 8-Entry
128
Store
Add
256
Fill
Coherent Memory Interconnect

Pipe 0 = ALU/Shift/Load/Store

Pipe 1 = ALU/Shift/MUL/COP2/Branch/Jump

- Thread-dedicated resources = very deterministic CPU performance
- Highly-associative L1 caches = equivalent miss rate to much larger caches

Achieving a 3 cycle load to use latency



LOAD	I	R	A	D1	D2	W			
ADD		I	R	A	D1	D2	W		
SUB			I	R	A	D1	D2	W	
NOR				I	R	A	D1	D2	W

Compiler must find ILP (load delay slots)

- Higher the dispatch width, the more instruction slots to fill
- Custom Circuit techniques
  - Not everything has to happen on a single cycle boundary
  - Need to do special timing analysis
  - Deterministic delay allows for optimal placement of registers/latches
  - Ability to build efficient, high speed fully associative structures
  - Usual other tricks (…)
- Optimize wires, logic for speed
- Have a simple load instruction
  - Alignment, simple address calculation, fewer exceptional conditions

Scale Out Data Center Processor

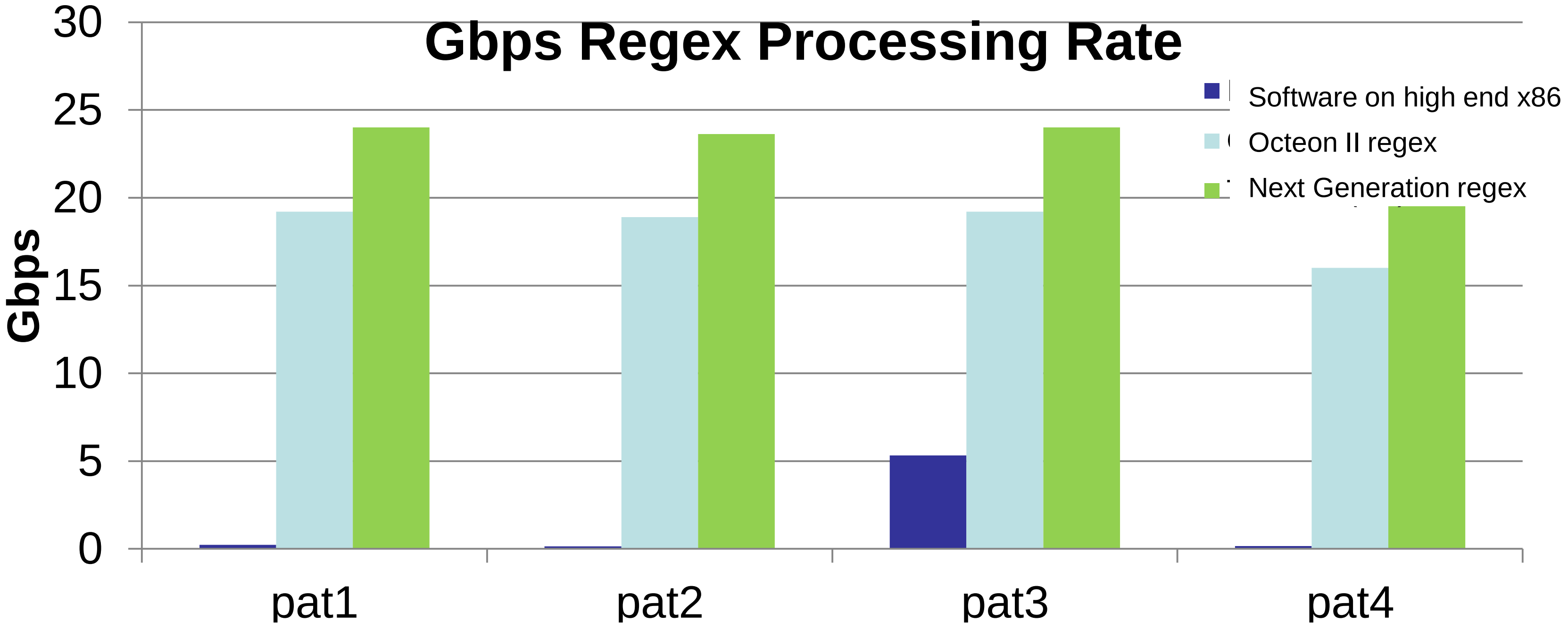
	Networking Processor	Data Center Processor	Remarks
Highly Threaded	✓	✓	
SSO (Scheduling Sync Unit)	✓		
Compression Accelerator	✓	✓	
Crypto Acceleration	✓	1/2	Focus is not on packet processing
Input Packet Parsing	✓	1/2	Data center more homogenous
Output queuing	✓	1/2	Data center more homogenous
High Bandwidth Networking	✓	✓	
Low Latency Networking		✓	
Regular Expression Engine	✓	✓	Can be repurposed
Integrated High Speed Network	✓	✓	Reduce components, improve reliability, lower power
Integrated Storage		✓	Rotating Media, SSD

Example: RegEx acceleration

- In Octeon, RegEx hardware used for packet sniffing
  - Intrusion detection
  - Virus detection
  - Packet classification
- In Data Center, Regex Hardware can be used
  - To parse text data – unstructured and semi structured data
    - Find ZIP codes, phone numbers, name, address
    - Search machine logs (error detection, site visit statistics)
  - Works well when setup time is small compared to run time : streaming bulk data!

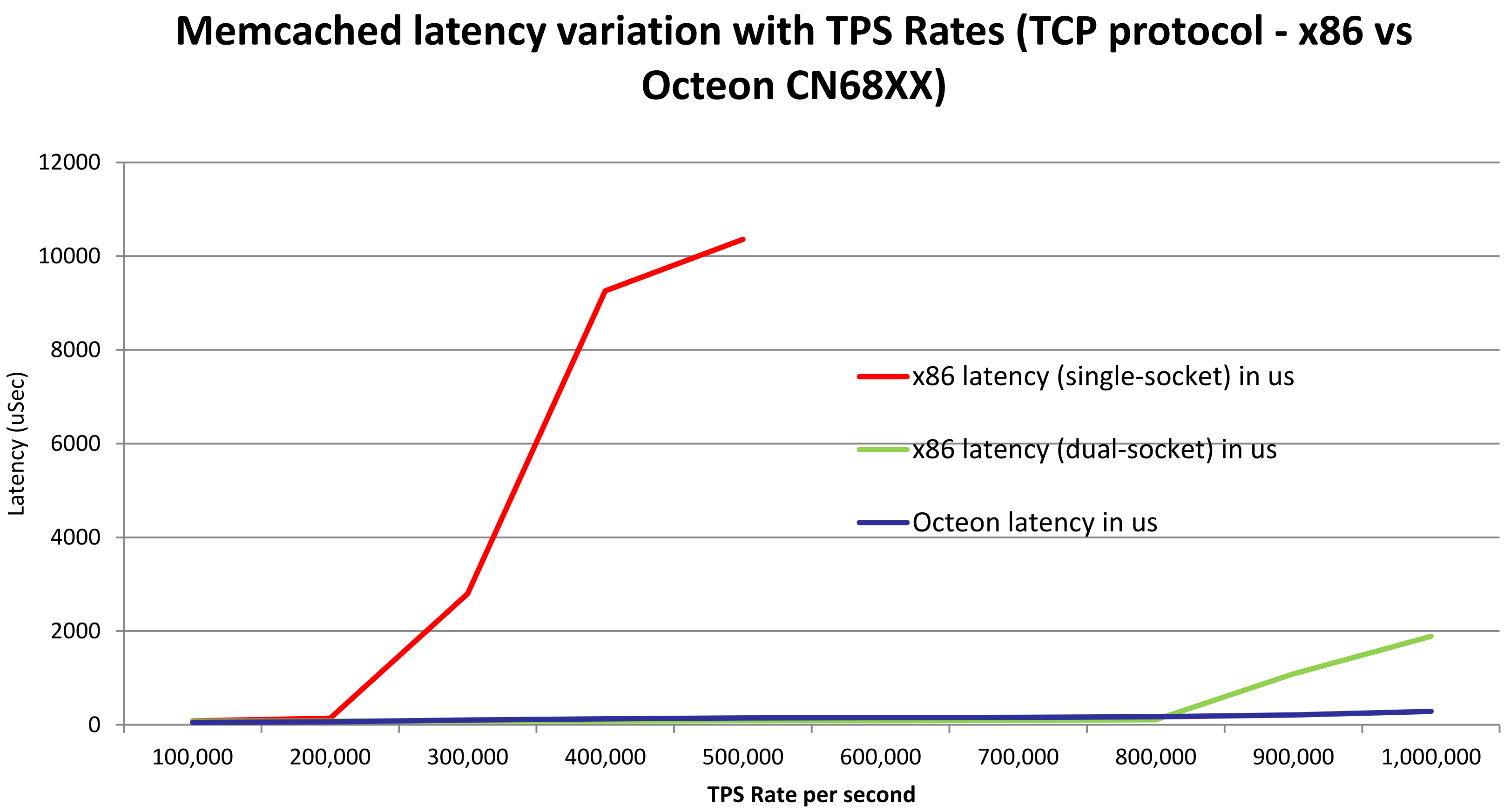
Example: Big Data Text Search (2)

- Text search
  - Submit precompiled regex pattern to regex engine
  - Search for 1 of 4 different patterns (e.g. url, email, date)
- Next Generation is 4.5 to 150 x faster than software only solution



benchmark from http://lh3lh3.users.sourceforge.net/reb.shtml

Memcached Latency Profile



- Single socket OcteonII compares well with dual socket Xeon X5-2690
- 2.9 GHz Xeon versus 1.5 GHz Octeon
- ***More cores is goodness***

Cavium Processors

- Networking Market
  - OCTEON Family
  - Well suited to area efficient cores (lots of aggregate processings)
  - Well suited to purpose built accelerators
- Scale Out Data Center
  - Well suited to area efficient cores
  - Well suited to purpose build accelerators
- SHAMELSS PLUG #1: ***It takes a lot of talented people – we are hiring! (bchin@cavium.com)***

Benchmarking

- How do we measure performance on these new kinds of applications?
  - Need to develop better metrics
  - System performance
    - Data center network
    - Disk I/O
    - OS software
- How do we measure agility
  - Configuration, maintainability
  - Elasticity
- Challenging problem
  - Very dynamic space
    - YARN, HIVE, PIG, Hadoop, Storm, Spark, Mahout, R,Presto, Drill, Scalding, Summingbird, Thrift, Impala, Parquet, SCUBA, Kafka,Cobbler, Chef
- SHAM***ELESS PLUG #2***
  - ***Talk to me about it.; eembc.org***
  - ***markus.levy@eembc.org; bchin@cavium.com***

Questions?

Poll – how many engineers does it take?

- (Core) Microprocessor design team (simple core)
  - Logic designers
  - Circuit and implementation
  - Verification
  - Physical design
  - Validation
  - Subtotal:
- Rest of chip
  - Logic designers
  - Circuit and implementation
  - Verification
  - Physical design
  - Validation
  - Subtotal:
- Total: ???

Poll – how many engineers does it take?

IMHO *[rough answers – based on experience at Sun, MIPS, QED, PMC, Cavium, etc.]:*

- (Core) Microprocessor design team (simple core)
  - Logic designers (~6) – FP Unit, Integer Unit, Load/Store, Instruction Fetch
  - Circuit and implementation (~6) – custom circuits
  - Verification (~6) – rule of thumb – 1 to 2 verif for each RTL
  - Physical design (~10)
  - Validation (~5)
  - Subtotal: ~30-40
- Rest of chip
  - Logic designers (~20)
  - Circuit and implementation (~10)
  - Verification (~20)
  - Physical design (~10)
  - Validation (~5)
  - Subtotal: 60-70
- Total: >100