PENSANDO

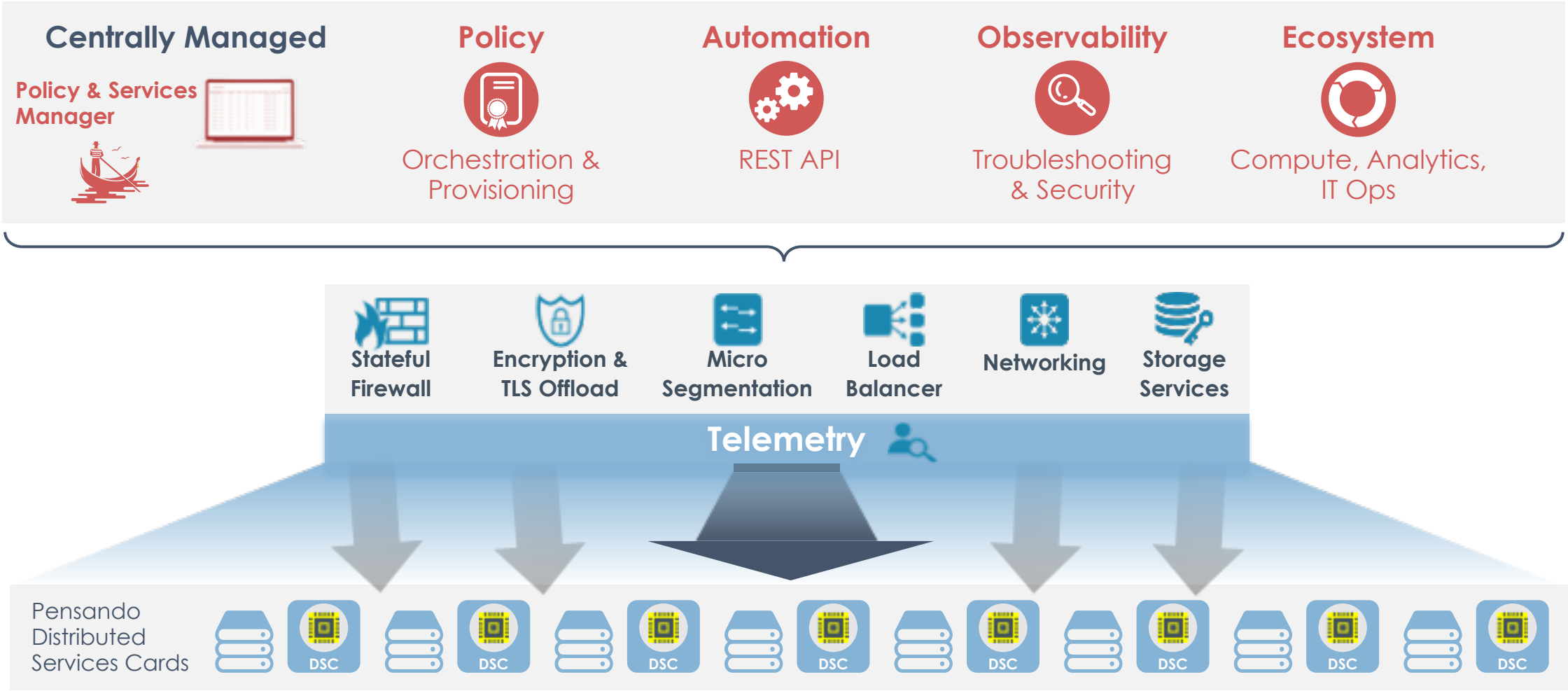
Distributed Services Architecture

Francis Matus

VP, Engineering

August 18, 2020

Pensando Distributed Services Architecture



No service stitching required: All functions are present everywhere at cloud scale

PENSANDO

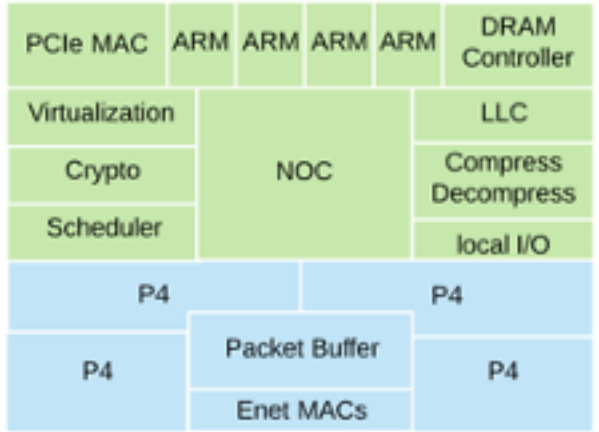
Pensando Chip Architecture

Networking Path (blue blocks in diagram)

- P4 pipelines process every packet entering and leaving the device
- Packet buffer switches packets between P4 pipelines and MACs
- P4 controlled DMA engines bridge the packet and memory domains

SOC Path (green blocks in diagram)

- NOC connects P4 with offload engines, ARM, PCIe, and DRAM
- Storage and crypto offloads operate on hardware queues controlled by P4, ARM, or host CPU
- P4 DMA transactions are IO coherent to ARM



P4 Pipeline Design

Pipeline starts with a Programmable parser that populates Packet Header Vector (PHV)

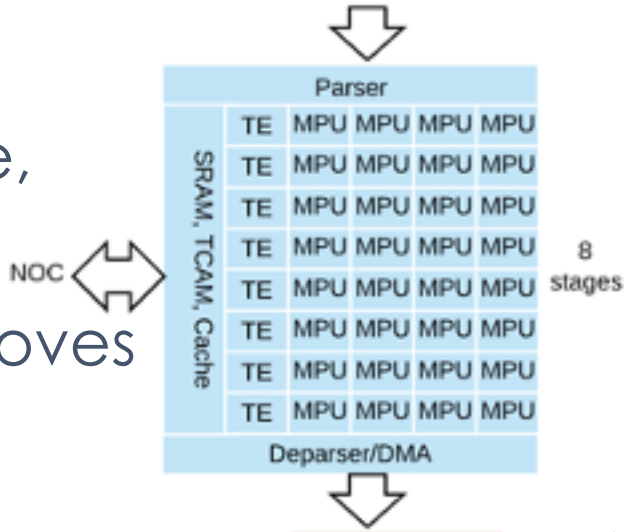
- PHV carries header fields, commands, and packet metadata
- PHVs are variable in length, up to 8 kbits

Up to 8 stages apply P4 match-action tables

P4 pipelines have local SRAM, TCAM, and DRAM cache resources to support table lookups

Pipeline ends with a deparser, using the PHV to remove, insert and/or rewrite the packet contents

If the PHV contains DMA commands, a local engine moves headers and payload to/from DRAM



PENSANDO

P4 Stage Design

Table Engine builds lookup keys up to 2048-bits wide

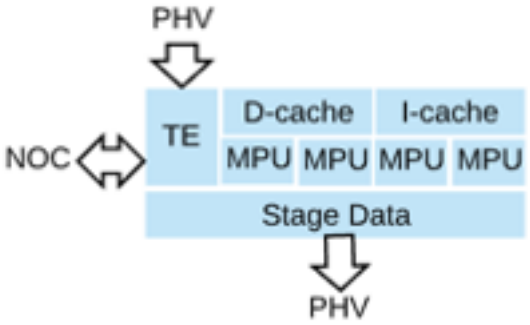
- Hash, TCAM, and direct index tables supported
- Multiple, non-dependent table lookups can be issued per packet
- Multiple pending table reads tolerates DRAM latency
- "locked" tables used to support single flow atomic table updates

Table results distributed to 4 Match Processing Units (MPUs)

- MPUs execute run-to-completion program associated with table
- MPU programs update the PHV, tables, and other data structures

Stage Data logic holds and updates PHVs from all MPUs

- PHVs graduate in-order to next stage in pipeline



Match Processing Unit (MPU) Design

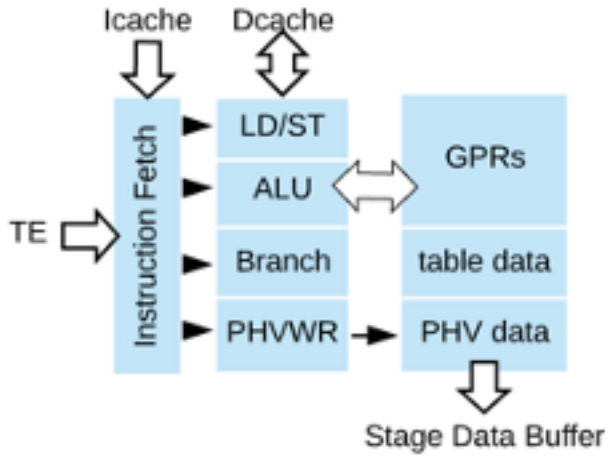
Table result and associated entry PC launch MPU programs

Domain specific Instruction Set Architecture

- Optimizes bit field manipulation, header field updates
- General purpose ALU, branch, logical, data movement, and control
- Specialized instructions for common protocol and queue operations

Code snippet example from TCP load balancer

```
seq        c1, p.tcp_valid, 1   // set c1 for TCP packets
phvwr.c1   p.daddr, d.lb_addr   // update daddr from table
tbladd.c1  d.byte_cnt, k.pkt_sz // update table byte count
```



Central Packet Buffer

Connects P4 pipelines and ethernet MACs

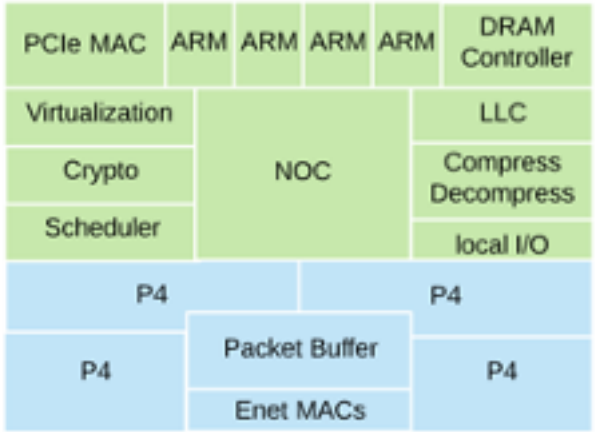
Shared memory switch

Data Center Ethernet Enhanced Transmission Selection (ETS) scheduler at outputs

Multicast replication and port mirroring

Priority-based Flow Control (PFC) absorption buffers

Packet burst overflow to DRAM, per COS



PCIe Virtualization Service

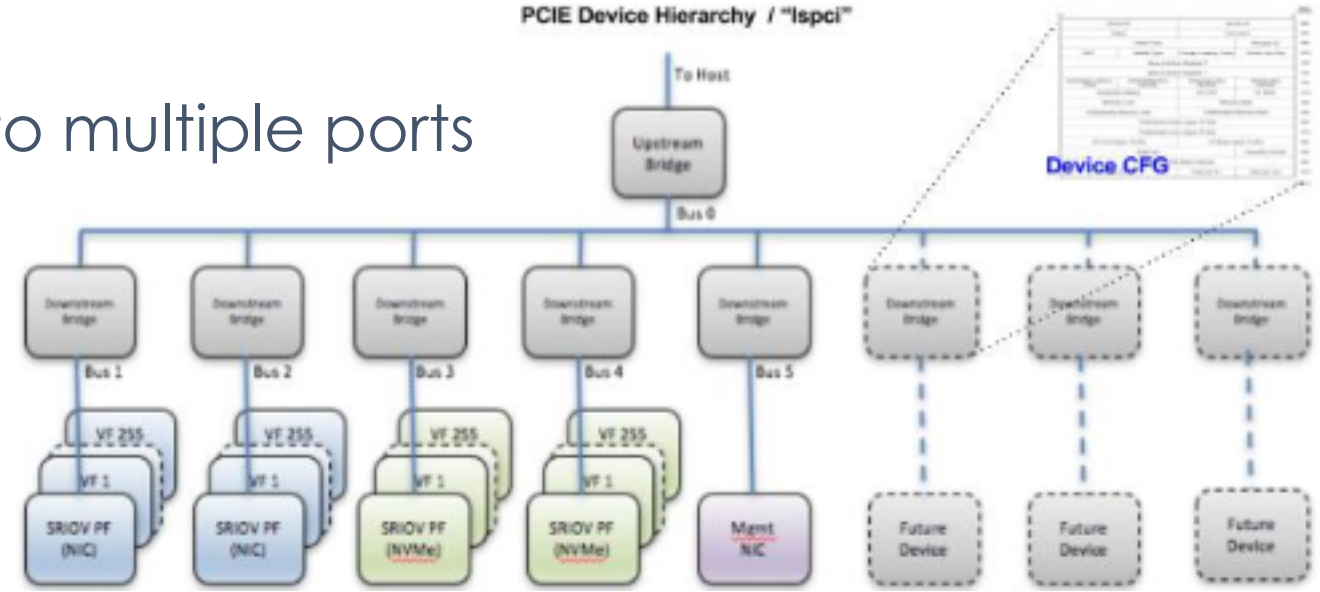
Chip presents full PCIe device topology, including PCIe switches, devices, config space, BARs, & resources:

- Ethernet NICs & RDMA devices
- nVME block storage devices
- Storage and security offload devices
- Future devices

PCIe lanes can be bifurcated into multiple ports

- Root Complex support

PCIe multi-host support



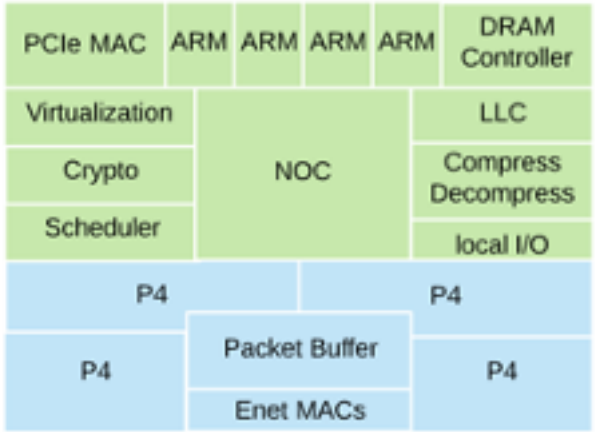
SOC, NOC, & Hardware Queues

NOC connects P4 DMA, offload engines, ARM, PCIe, and DRAM in both coherent and non-coherent domains

ARM CPUs run SMP Linux and interface to P4 with either a netDev or DPDK interface

Doorbells for 16 million hardware queues are mapped to host or ARM processes and initiate multi-level scheduler to inject P4 pipeline tokens

Ethernet and DPDK device drivers available for multiple host OSes



Root of Trust, Cryptography

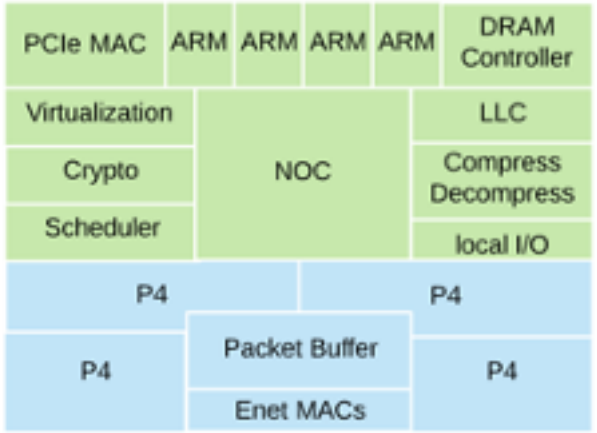
Root of Trust based on Physical Unclonable Function (PUF)

Secure Boot loader and OS authentication before execution

P4 programs terminate and proxy secure protocols (IPsec, TCP/TLS), ARM programs also have access to offload engines

Security Offload Engines

- Secure connections: Deterministic and True Random Number Generator and Public Key Exchange
- Cryptography offloads: AES-GCM, AES-XTS, AES-CBC, AES-CCM, ChaChaPoly, HMAC, SHA3-512, and SHA2-256



Storage Offloads

Compress/Decompress Offload Engines

- LZRW1A, GZIP and Deflate algorithms at 100 Gbps

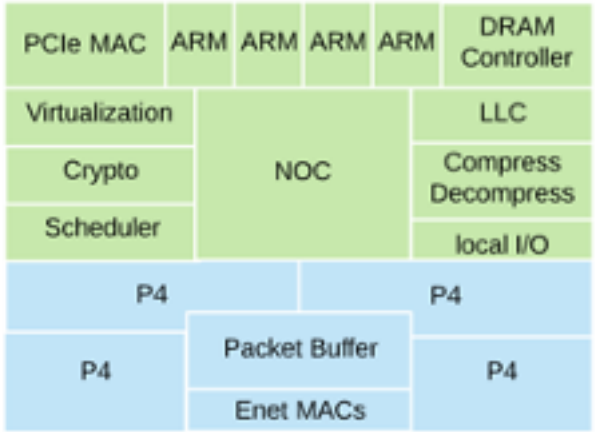
Reed-Solomon Erasure coding Engines

- Up to 12 data and 4 parity blocks operate at 100 Gbps

Data integrity engines operate at 200 Gbps

- CRC64, CRC32C, Adler-32, and M-Adler32.

De-duplication support based on SHA2 and SHA3 engines



P4-16 Compiler

Permissively licensed, domain specific P4 language

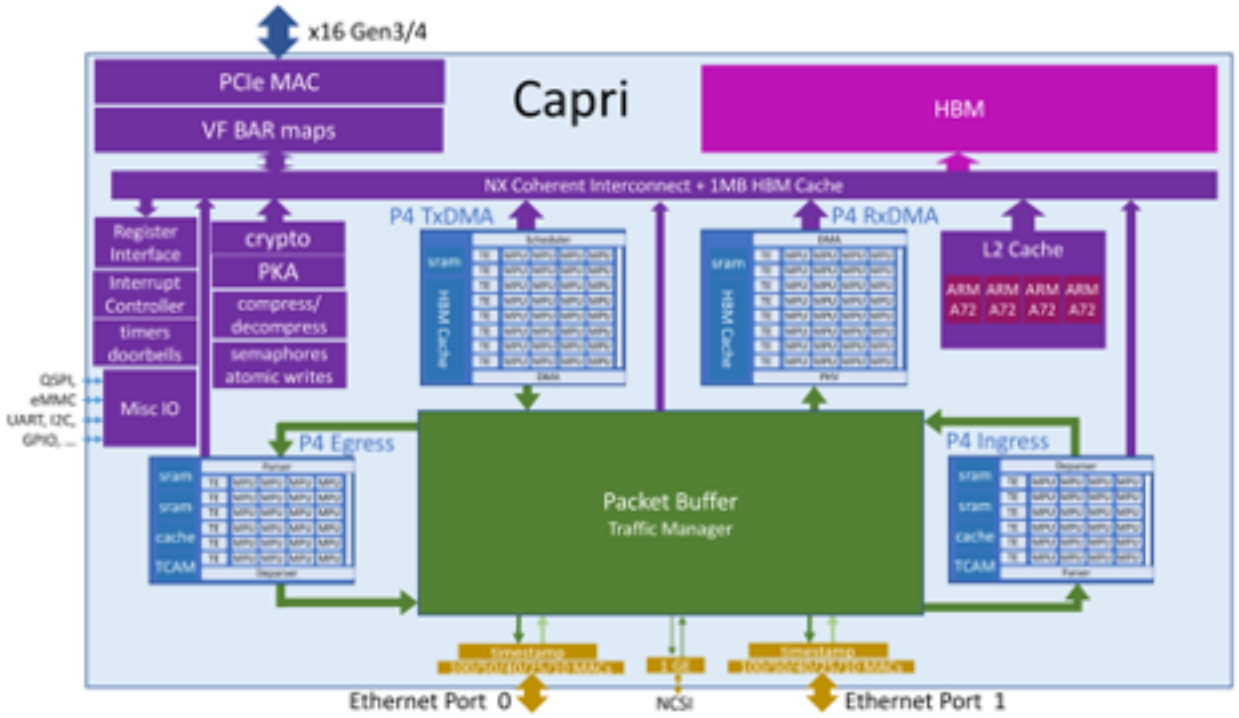
Pensando LLVM-based compiler generates

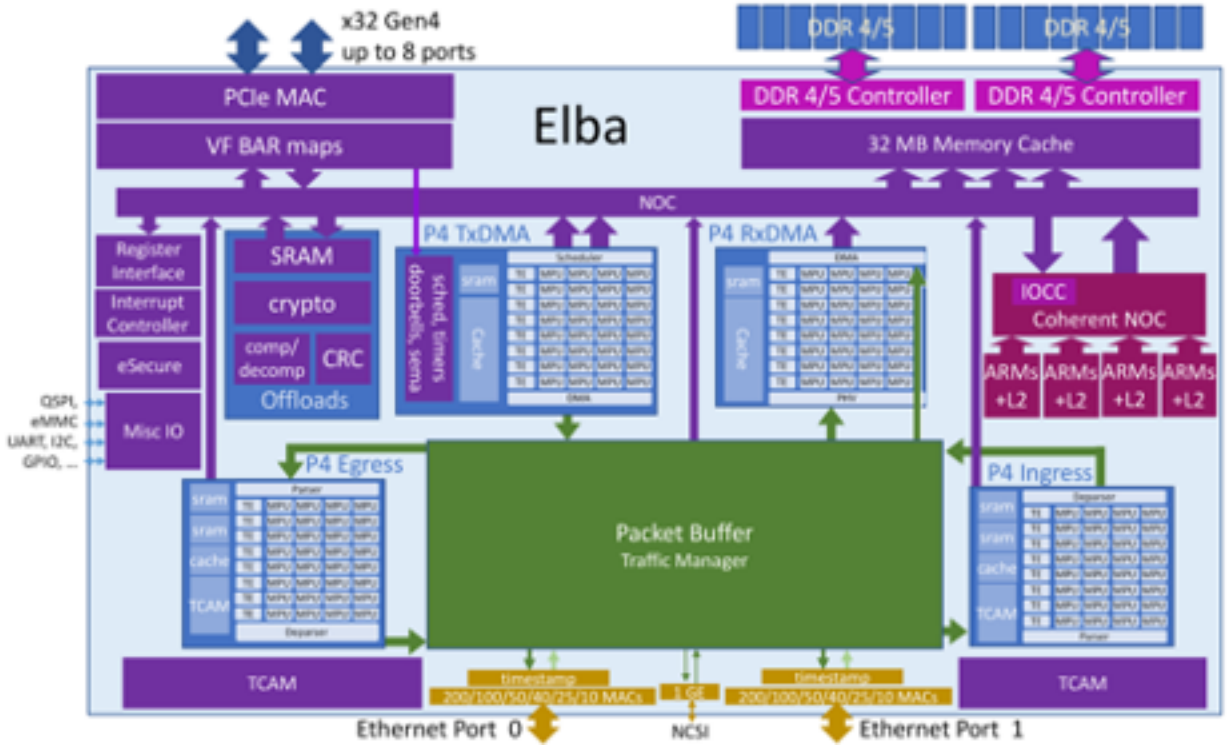
- Packet parser command sequences
- Table configuration, allocation and memory placement
- MPU action instruction sequences

For packet forwarding pipeline, compiler generated code achieves 85% performance of hand coded assembly

As hardware evolves and improves features, existing P4 source code can be quickly ported to future implementations

Capri (16nm) and Elba (7nm) Block Diagram





PENSANDO

Chip Performance, Implementations

Chip Performance
(SDN bump-in-the-wire application)

Feature	**Capri**	**Elba**
Latency	3usec	3usec
Jitter	3.5nsec	3.5nsec
Packets per Second	40 Million	80 Million
Connections per Second	1M	2.5M

Chip	**Capri**	**Elba**
Technology	16 nm	7 nm
Prototypes	Sept '18	June '20
Network	NRZ 25G 2 * 100G or 4 * 50,25,10	PAM4 50G 2 * 200G or 4 *100,50,25,10
PCIe Gen4	16 lanes	32 lanes
Memory	2.5D HBM 4-8 GB	dual DDR4/5 8-64 GB
P4 cores Frequency	112 MPUs 830 MHz	144 MPUs 1.5 GHz
ARM cores Frequency	4 * A-72 2.2GHz	16 * A-72 2.8GHz
L2 cache LL cache	2 MB 1 MB	8 MB 32 MB
power	< 30W	< 25W@100GE < 50W@200GE

THANK YOU