FORWARDING HARDWARE IN NETWORKING

Xantaro
SERVICE INTEGRATION

WOAMI

- Tobias Heister
- Solutions Architect
  - Technical Pre-Sales
  - Vendor MGMT for Focus Vendors
  - Solutions Engineering and Partner Evaluation
  - XT3LAB

- 3y Deutsche Telekom – Carrier and DC
- 8y Host Europe (now GoDaddy) - Hosting Provider, Network centric Roles (Engineering, Managed Services)
- 6y Xantaro – SOLAR

AGENDA

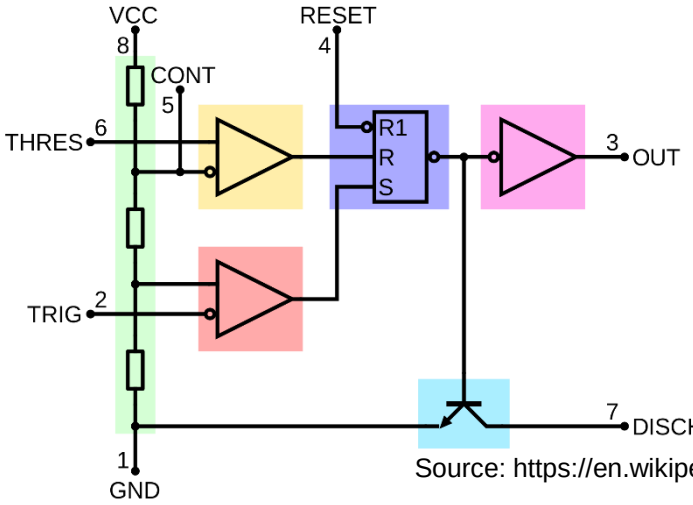
- Definition and Wording (ASIC, FPGA, NPU)
- Why do we need ASICs?
- Run to Completion vs Pipeline Architecture
- Fixed vs. Progammable Pipelines
- Metrics for ASICs (Troughput, Interfaces, Buffers)
- Merchant Silicon vs. Proprietary
- When to use Which ASIC?
- ASIC Landscape (vendor and product mapping)

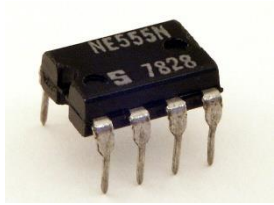
DISCLAIMER

- Most of the Info is hard to find
  - Sometimes multiple sources contradict each other
  - a lot is hidden behind NDAs and cannot be shared
- I am mostly talking about things i have experience with or have touched
  - There are more vendors and way more variants of chips out there
- If you know better, please let me know to update and enhance the info!

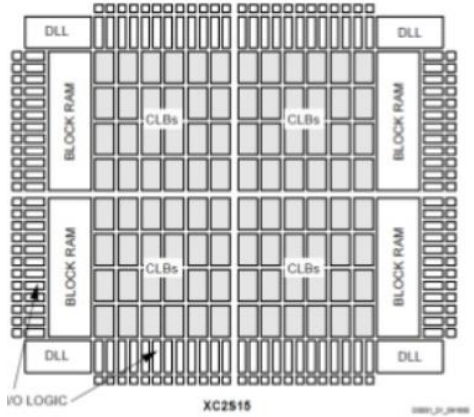
DEFINITION AND WORDING

- Integrated Circuit (IC)
- Application Specific Integrated Circuit (ASIC)
  - very fast, optimized in sized, does „one“ job
  - built once, cannot be changed during lifetime
- Field Programmable Gate Array (FPGA)
  - Configurable logic blocks (CLB)
  - Design based on Software (e.g. via a HDL)
  - Slower than ASICs, „larger“, more versatile
- Micro/Graphics/Network Processor Units (MPU, GPU, NPU)
  - Specialized electronic circuits
  - Need Software/Programming to perform tasks





Source: https://en.wikipedia.org/wiki/555_timer_IC



Source: http://webdocs.cs.ualberta.ca/~amaral/courses/329/webslides/TopicG-FPGAOrganization/img2.html

DEFINITION AND WORDING (CONTINUED)

- In Networking Forwarding Hardware (commonly called ASIC) typically means a combintion or hybrid of
  - actual ASICs and or blocks of ICs
  - NPUs using external memory
  - sometimes FPGA to support specific scenarios
  - general purpose (x86, ARM, MIPS) CPUs
  - all of the above

Network processor

From Wikipedia, the free encyclopedia
(Redirected from Network Processing Unit)

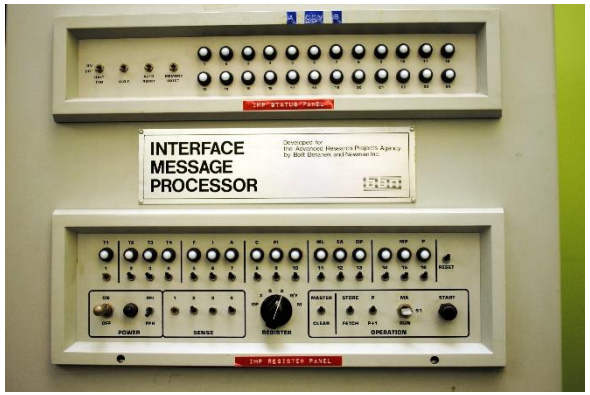
A **network processor** is an integrated circuit which has a feature set specifically targeted at the networking application domain.

Network processors are typically software programmable devices and would have generic characteristics similar to general purpose central processing units that are commonly used in many different types of equipment and products.

Source: https://en.wikipedia.org/wiki/Network_processor

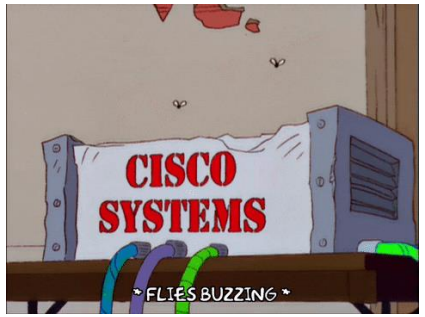
WHY DO WE NEED ASICS

- Truth and Myth
  - Software is Slow? Hardware is Fast?
- Even ASICs run „programms“
  - They are just pretty good at doing it very fast
  - They are not good at doing anything else
- It´s always a race between Software and Hardware
  - Software based forwarding is more versatile but slow (Cisco 7200)
  - Hardware based forwarding is faster (Juniper M40)
  - Until the next processor generation arrives
  - Until the next ASIC generation arrives
  - Repeat



Source: https://en.wikipedia.org/wiki/Interface_Message_Processor

Source: https://gblogs.cisco.com/ru/cisco_certification_history/



Source: https://media.giphy.com/media/3o6MbhpKReYen8pnCo/source.gif

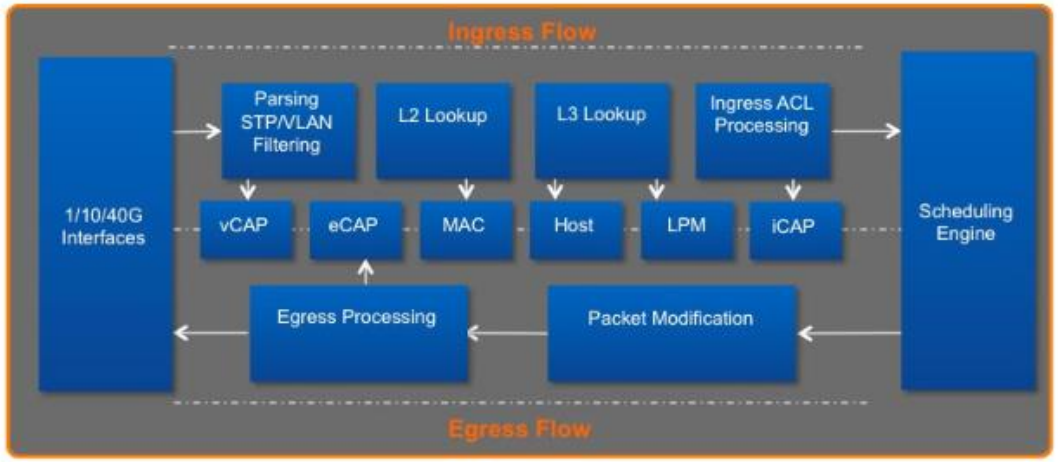
Source: https://www.townsendassets.com/wp-content/uploads/2010/11/Juniper_M40.jpg

DPDK
DATA PLANE DEVELOPMENT KIT

Xantaro SERVICE INTEGRATION

DIFFERENCES IN NPU

- Pipeline based Architectures
  - a number of steps in a fixed order
  - each step is limited to doing specific things
    - only specific combinations of features may be possible
    - sometimes recirculation is possible
  - fixed/known runtime trough pipe for each packet
  - „cheap“ but less versatile

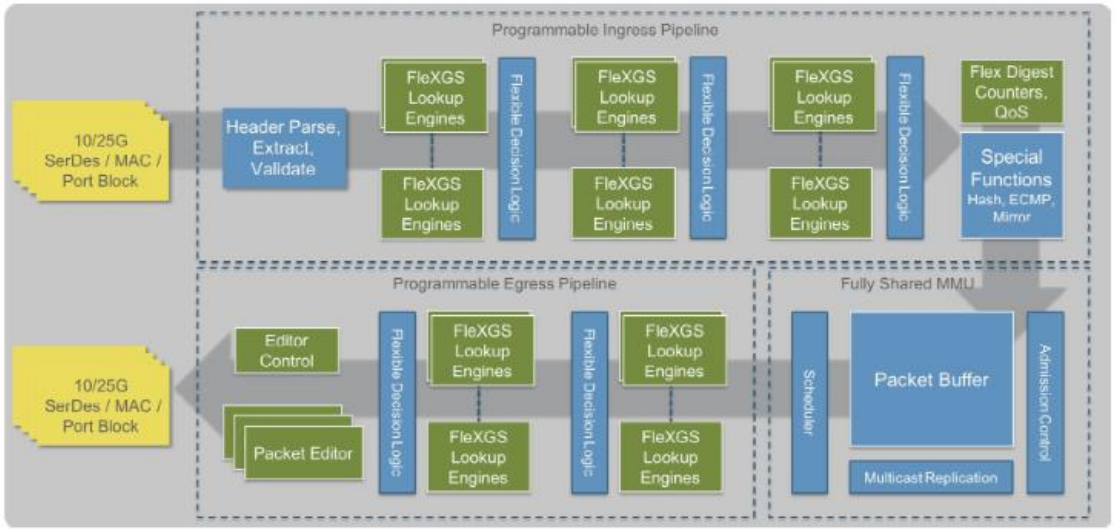


Source: https://www.arista.com/assets/data/pdf/Whitepapers/Arista_7050X_Switch_Architecture.pdf

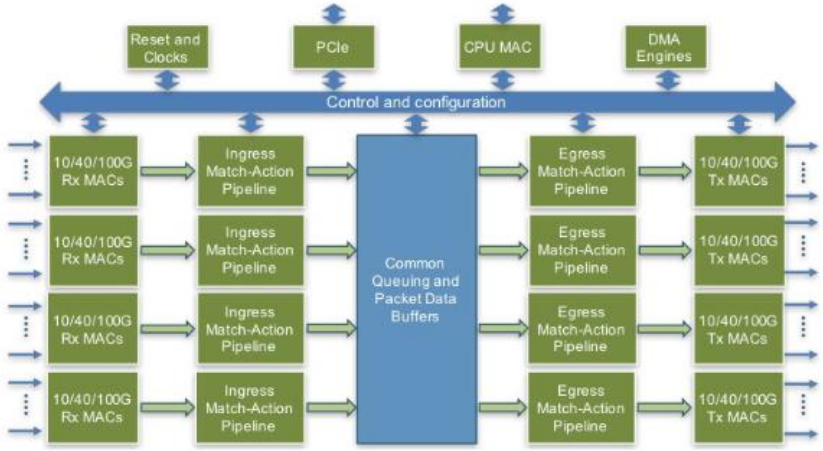
- Run to Completion (RTC)
  - more like a CPU - runs a programm which may branch and loop
    - undetermistic runtime (typically tries to avoid infinity loops etc.)
    - runtime may vary per packet
  - can (often) learn new features
  - „expensive“ but flexible

FIXED VS. PROGAMMABLE PIPELINE

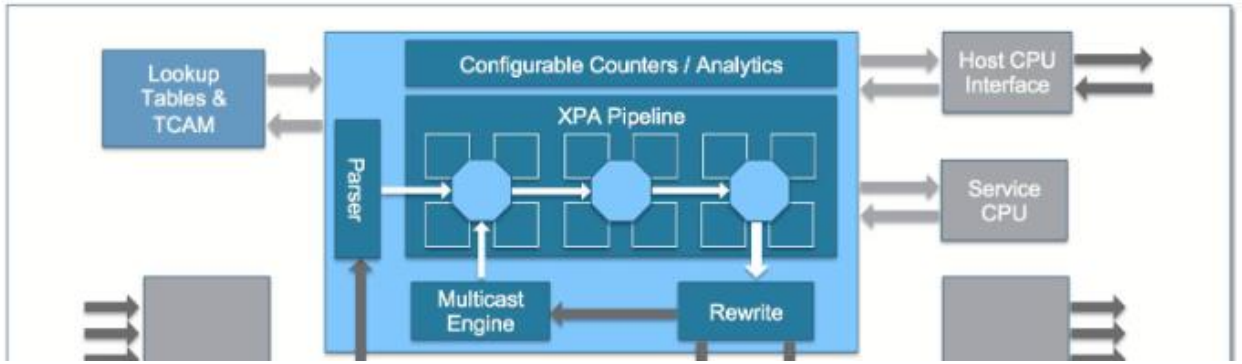
- fixed Pipelines can be limiting
- more and more chips offer some programmabilty
- we can change what happens at which stage
- we (typically) can not change the number of stages
- the lines between pipeline and RTC are bluring



Source: https://www.arista.com/assets/data/pdf/Whitepapers/7050X3_Architecture_WP.pdf



Source: https://www.arista.com/assets/data/pdf/Whitepapers/7170_White_Paper.pdf



Source: https://www.arista.com/assets/data/pdf/Whitepapers/7160SwitchArchitecture_WP.pdf

METRICS FOR ASICS

- Bandwidth in Tbit/s
  - ideal conditions (larger packets, „low“ pps)
  - not nessecarily using all features
- Amount and Speed of SerDes (Pins)
  - Front Facing
  - Fabric Facing
  - As of today often 10, 25G per SerDes (56, 112G are coming)
  - Supported Interfaces speeds (1, 10, 25, 40, 50, 100, 400)
    - with/without gearboxes
- Buffer
  - On Chip/Off Chip
  - Size - Mega Byte vs. Giga Byte
  - shallow/medium vs. deep buffer

BCM88670

StrataDNX® Jericho switch series

OVERVIEW SPECIFICATIONS

Specification	Value
Lifecycle	Active
Distrib. Inventory	No
SerDes / GPHY	24 x 25G + 48 x 10G
I/O Bandwidth (Gbps)	1.44
Bandwidth UOM	Tb/s
Bandwidth	1.44Tb/s
Fabric	36 x 25G

Source: https://www.broadcom.com/products/ethernet-connectivity/switching/stratadnx/bcm88670

METRICS FOR ASICS

- Table Sizes (FIB)
  - Internal or external Lookup Tables
    - Internal are „fast“, external might be „slow“
  - type of memory
    - TCAM – fast and expensive
    - $RAM – slow and cheap
  - Fixed or dynamic
    - some chips have fixed slices
    - some chips have dynamic allocation
      - Fully dynamic and decided as needed (more complex, typically in RTC Systems)
      - specific partition schemes (Unified Forwarding Table – UFT on Trident is an example)
    - some vendors use tricks to squeeze more out of the available space

Table 5: Arista 7050X3 UFT modes

UFT Mode	0	1	2 Default	3	4
MAC Addresses	288K	224K	160K	96K	32K
IPv4 Host Routes	16K	80K	144K	168K	16K
IPv4 Multicast (S,G)	8K	40K	72K	104K	8K
IPv6 Host Routes	8K	40K	72K	104K	8K

Source: https://www.arista.com/assets/data/pdf/Whitepapers/7050X3_Architecture_WP.pdf

Table 6: Arista 7050X3 ALPM mode

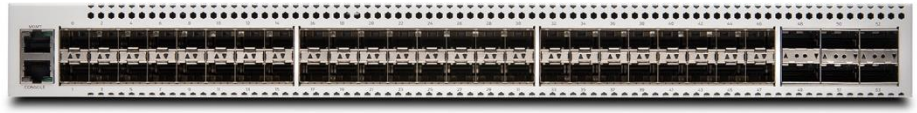
LPM Table Mode	ALPM	1	2	3	4
IPv4 LPM Routes	384K	32K	32K	32K	32K
IPv6 LPM Routes Unicast (Prefix Length <= /64)	192K	12K	8K	4K	--
IPv6 LPM Routes Unicast (any prefix length)	40K	2K	4K	6K	8K

MERCHANT SILICON VS. PROPRIETARY

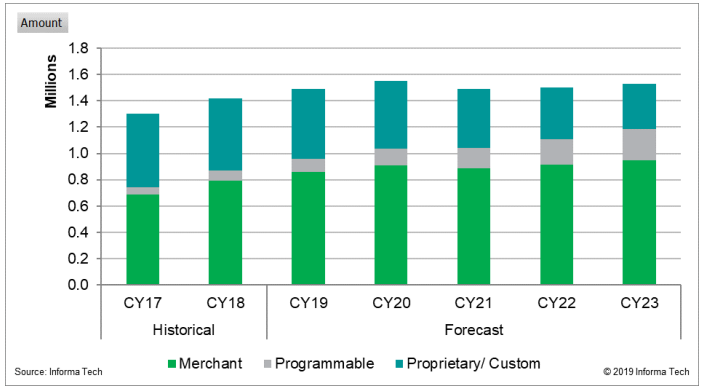
- Proprietary
  - In the beginning all Chips were vendor specific (Cisco, Juniper and others before)
  - IP of the Chip as well as the SDK are kept in House
  - Chip will only be used in products of one vendor
  - Most Big vendors are still doing this as of today
  - Feature Rich, Versatile, Purpose built, long history, expensive (all R&D is in House)
- Merchant Silicon
  - Chip and SDK is sold on the market, Sometimes Chips Specs are shared and SDK can be bypassed
  - Other vendors design products around the chip
  - Not that feature rich (but evolving very fast), can be cheaper as R&D Cost is shard between multiple parties
  - Companies like Marvel, Broadcom and others design and sell chips but have no own products
  - Some big vendors are opening up now

MERCHANT SILICON ON THE RISE

- All big vendors have products using Merchant Silicon
  - it started in switching
  - continued in routing and Metro/MPLS
  - extended into the optical world
- Merchant Silicon is driving innovation
  - „fastest“ Chip typically is merchant silicon
  - „cheapest“ Chip typically is merchant silicon
  - good enough for >95% of all use cases



Source: https://www.juniper.net/assets/img/products/image-library/ocx1100/ocx1100-front-high.jpg



Source: https://elegantnetwork.github.io/posts/A-Summary-of-Network-ASICs/

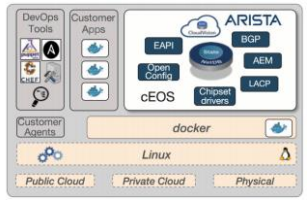
Source: https://5.imimg.com/data5/EH/YE/MY-59394742/500-500x500.jpg

- Proprietary still relevant
  - Earliest adoption of features as implementation is easier (especially in RTC Architecture)

MERCHANT SILICON AND WHITEBOX

- Whitebox
  - as you can buy the Chips somewhere else, why not buy the complete switch
  - $emerging vendors (most of the time the companies doing it anyway) now offer devices based on merchant silicon
    - in house design or designs based on or published as OCP documents
    - some of the big vendors just buy these boxes and port their NOS on it
    - the $cloud Guys do it as well

- Software
  - not only using chip vendor SDK to control the Chip
  - build an own NOS and Control Plane
  - Commercial and Open Source Options available



WHEN TO USE WHICH ASIC?

- Most of the time somebody else decided for you
  - vendors will build a box using a specific chip and market it to be used to a specific use case
  - NOS and other Software might be artificially limited to suit that use case

- if vendors offer more than one box with similiar look and feel, it might be important to understand the difference

Sources: https://www.juniper.net/us/en/company/press-center/images/

WHEN TO USE WHICH ASIC – BROADCOM (ALL PIPELINE BASED)

- Strata XGS (Missile/Weapon)
  - Trident
    - Trident2 1,2 Tbit/s, 12MB Buffer, Edge Features
    - Trident2+ 1,2 Tbit/s, 16MB Buffer, Edge Features
    - Trident3 3,2 Tbit/s, 32MB Buffer, Edge Features
    - Trident4 12,8Tbit/s, 132MB Buffer, Edge Features
  - Tomahawk
    - TH 3,2 Tbit/s, 16MB Buffer, Spine Features
    - TH2(+) 6,4 Tbit/s, 42MB Buffer, Spine Features
    - TH3 12,8 Tbit/s, 64MB Buffer, Spine Features
    - (TH4) 25,6 Tbit/s, XXMB Buffer

- Strata DNX (Sand, Desert Cities)
  - Jericho
    - 720 Gbit/s, 9GB Buffer, Metro and IP Features
    - 1M FIB Entries
  - Jericho+
    - 900 Gbit/s, 9GB Buffer, Metro and IP Features
    - 2M FIB Entries
  - Jericho2
    - 10 Tbit/s, 8GB Buffer, Metro and IP Features
    - 2,5M FIB Entries
  - Qumran (AX, UX, MX) – fabricless Jericho
    - up to 800Gbit/s, 3GB Buffer
    - 1M FIB Entries

WHEN TO USE WHICH ASIC - JUNIPER

- Trio (Run to Completion)
  - many different generations (starting in 2007)
  - used in MX series
    - most recent Generation called ZT
      - 500 Gbit/s, 10M FIB entries
    - Extremely feature rich
      - If you can name a feature it probably can support it
    - learned many new tricks over the years
      - VXLAN, SR, Hyper-Mode, Flex Filter and many more
    - 1, 10, 40, 100, 400G (now even 25G)

- Express (Programmable Pipeline)
  - Many different generations (starting in 2012)
  - Started on PTX later branched out into QFX10k
    - most recent generation called BT
      - 3,6 Tbit/s
    - Lean feature set, but growing constantly
    - Learned many new tricks over the years
      - from LSR to Full Table, VXLAN, SR and many more
    - 1, 10, 40, 100, 400G (no 25/50G)

WHEN TO USE WHICH ASIC - OTHERS

- Nokia
  - FP4 (supposed to be Run to Completion)
    - 3TBit/s
      - clear chanel 1TBit/s, flexible smart filtering
      - very feature rich

- Cisco
  - Silicon One Q100 (programmable pipeline?)
    - 10.8 Tbit/s
      - probably lean feature set, might be sold to third parties
      - P4 capabale
      - „global Routing Scale“
      - deep buffer

WHEN TO USE WHICH ASIC - OTHERS

- Barefoot (Intel)
  - (fully programmable pipeline)
  - P4 centric, Very versatile use cases (depend on P4 code)
    - Tofino, 6,4 Tbit/s
    - Tofino 2, 12,8 Tbit/s

- Innovium
  - (programmable piple)
    - TERRALYNX 7 - 12,8 Tbit/s
    - TERRALYNX 8 - 25,6 Tbit/s

- Cavium (Marvell)
  - XP80 (fully programmable pipeline)
    - 3,2 Tbit/s

- Mellanox (Nvidia)
  - (programmable pipeline)
    - Spectrum 2 - 6,4 Tbit/s, 42MB
    - Spectrum 3 - 12,8Tbit/s, 64MB

ASIC LANDSCAPE (VENDOR AND PRODUCT MAPPING)

- Juniper
  - QFX 5000
    - QFX51XX — BCM Trident
      - QFX5100 — BCM Trident2
      - QFX5110 — BCM Trident 2+
      - QFX5120 — BCM Trident3
      - QFX5130 — BCM Trident4
    - QFX52XX — BCM Tomahawk
      - QFX5200-32C — BCM TH
      - QFX5200-48Y — BCM TH+
      - QFX5210-64C — BCM TH2
      - QFX5220-32D — BCM TH3

- Juniper
  - QFX10000
    - QFX10002 — Juniper Express PE 500G
      - 36Q — 3 PFE
      - 72Q — 6 PFE
      - 60C — 12 PFE
    - QFX10008/100016 — Juniper Express PE 500G

ASIC LANDSCAPE (VENDOR AND PRODUCT MAPPING)

- Juniper
  - ACX Series
    - ACX1000-4000
      - BCM Enduro
    - ACX5048/96
      - BCM Trident 2
    - ACX710/ACX5448
      - BCM Qumran

- Juniper
  - EX Series
    - EX2xxx, 3XXX + 4300 (including MP)
      - low-end BCM
    - EX4600
      - BCM Trident2
    - EX4650
      - BCM Trident3
    - EX9200 (MX240-960 in disguise)
      - Juniper Trio based Line Cards (MPC4 – MPC7)
    - EX9250 (MX240/10003 in disguise)
      - Juniper Trio based EA Generation

ASIC LANDSCAPE (VENDOR AND PRODUCT MAPPING)

- Juniper
  - MX Series
    - MX80/104 — Trio 40G/80G
    - MX204 — Trio 400GE
    - MX150 — virtual Trio (X86)
    - MX240-960
      - MPC1E — Trio 40G per PFE, 1 PFE
      - 16x10MPC — Trio 40G per PFE, 4 PFE
      - MPC2/3E-NG — Trio 130G per PFE, 1 PFE
      - MPC4E — Trio 130G per PFE, 2 PFE
      - MPC5E — Trio 120G per PFE, 2 PFE
      - MPC7E — Trio 240G per PFE, 2 PFE
      - MPC10E — Trio 500G per PFE, 2-3 PFE
- Juniper
  - MX2000
    - MPC6E — Trio 130G per PFE, 4 PFE
    - MPC8E — Trio 240G per PFE, 4 PFE
    - MPC9E — Trio 400G per PFE, 4 PFE
    - MPC11E — Trio 500G per PFE, 8 PFE
  - MX10008/16
    - LC2101 — Trio 400G per PFE, 6 PFE
  - MX10003
    - LC2103 — Trio 400GE per PFE, 3 PFE

ASIC LANDSCAPE (VENDOR AND PRODUCT MAPPING)

- Arista
  - 7050X — BCM Trident
    - 7050X — BCM Trident 2
    - 7050X2 — BCM Trident 2+
    - 7050X3 — BCM Trident 3
  - 7060X — BCM Tomahawk
    - 7060X/7260X — BCM TH
    - 7060X2 — BCM TH+
    - 7060X3/7260X3 — BCM TH2
    - 7060X4 — BCM TH3

- Arista
  - 720XP — BCM Trident
    - 720XP — BCM Trident 3
  - 7280R — BCM DNX
    - 7280R — BCM Jericho
    - 7280R2 — BCM Jericho+
    - 7280R3 — BCM Jericho2
  - 7020R — BCM DNX
    - 7020TR — BCM Qumran AX
    - 7020SR — BCM Qumran AX

ASIC LANDSCAPE (VENDOR AND PRODUCT MAPPING)

- Arista
  - 7500 Chassis
    - 7500R Cards — BCM Jericho
    - 7500R2 Cards — BCM Jericho+
    - 7500R3 Cards — BCM Jericho2
  - 7800 Chassis
    - 7800R3 Cards — BCM Jericho2

- Arista
  - 7300X Chassis — BCM Trident 2
  - 7300X3 Chassis — BCM TH+
  - 7300X4 Chassis — BCM TH3
  - 7010 — BCM low end
  - 7170 — Barefoot Tofino
  - 7160 — Cavium XP80
  - 7130 — Metamako Aquisition

ASIC LANDSCAPE (VENDOR AND PRODUCT MAPPING)

- Cisco
  - ASR9000 + 9900
    - Some First Line Card generations were EZChip (now Nvidia)
    - NOTE: codenamed Trident and Tomahawk but have nothing to do with the BCM Chips of the same Name!
    - following Line Card generations have Cisco Chips
  - NCS5500
    - BCM DNX based (mainly Jericho+ and Jericho2)
  - ASR920
    - BCM DNX based (Qumran)

- Cisco
  - 8000 Series
    - Silicon One Q100

- Nokia
  - SR and XRS Series
    - Nokia FP3 + FP4
  - IXR Series
    - BCM DNX (Qumran to J2)
  - SAS Series
    - BCM low end

FURTHER READING

- Arista offers HW Architecture Guides for most platforms, they are pretty good and cover packet walkthroughs
- https://people.ucsc.edu/~warner/buffer.html offers a pretty good overview (although dated)
- Cisco Networker Breakout Sessions
  - e.g. ASR9K: https://www.ciscolive.com/c/dam/r/ciscolive/emea/docs/2020/pdf/BRKARC-2003.pdf
- Juniper Day One Guideds – Inside the MX 5G
  - https://www.juniper.net/documentation/en_US/day-one-books/DO_MX5G.pdf

DANKE FÜR DIE AUFMERKSAMKEIT