Stanford Guest Lecture
Seminar Course: Technologies in Finance
April 08, 2019

Mohan Kalkunte, Ph.D.
Vice President, Architecture & Technology
Core Switching Group
Broadcom Inc.
Design of a Switch Chip
Before you decide to build a chip

• Business Case
  • Market Segments
  • Major Customers
  • Key features of chip
  • Risks
  • ROI

• Market Segment
  • Hyperscale Data Center
    • e.g., Facebook, Microsoft, Google, Amazon
  • Enterprise Data Center
    • e.g., Large Fortune 500 company
Where in the Network

- Top of Rack (ToR)
- Leaf/Spine
- Aggregator etc.

Facebook Datacenter Architecture
Source: https://code.fb.com/data-center-engineering/f16-minipack/

© 2019 Broadcom. All Rights Reserved. The term “Broadcom” refers to Broadcom Inc.
Switch Silicon Design: A Highly Constrained Problem

Optimal switch silicon needs to meet or exceed on all these vectors

AND
Within the market window
Typical Data Center features

- High bandwidth
- Lower latency
- Large IP Routing
- Equal Cost Multi Path (ECMP)
- Hashing
- ACLs
- Monitoring
  - sFlow
  - Mirroring etc.
Constraint: Max Die Size limit

- Current hard-limit on silicon die size
  - 26mm x 33mm
  - dictated by reticle size
  - Practical size ~ 800 sqmm
    - Tight margin for error

http://www.silicon-edge.co.uk/j/index.php/resources/die-per-wafer

© 2019 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc.
Constraint: Cost

• One time cost – amortized over the product volume
  – Development cost
  – Mask costs

• Device cost
  – Die + package + test
  – Yield
    – Improves over time then flattens
    – Falls exponentially with size or complexity
  – Repair is a must for memories

• Memory is repairable
  – Row and column redundancy
  – Lower cost per sqmm for memory after repair

Source: anysilicon.com

Source: www.ee.ryerson.ca/~courses/coe838/lectures/SoC-IC-Basics.pdf
Constraint: IO Speed (Serdes speed)

• Tomahawk3 – 256 x 50G – 12.8Tbps

• Single device switch bandwidth keeping up with exponential increase

• Criteria
  – Reach
    – Copper Cables – Higher signal loss per unit distance
    – Optics: lower signal loss per unit distance
  – Cost / area

Source: ethernetalliance.org roadmap
Constraint: Power Dissipation

- Heat sink
- Heatsink with heatpipes
- Cold plate technology
- Immersion cooling

No redundancy for fan failure
- Fans have higher failure rate
Constraint: Process Geometry

Choice of Buffer Architecture

• Many buffer architectures are possible

• Which is the best choice?
  • Depends
EFFICIENT BUFFER ARCHITECTURE

• High burst absorption
  • Unused packet buffer available for transient congestion

• Fairness under congestion
  • Fair access to all ports and queues under heavy traffic load

• Avoid Starvation
  • Congested port should not starve uncongested ports

• Low frame loss
  • High zero-loss throughput performance

• Traffic Independent Performance
  • Buffer management with minimal tuning

• Scalable across multiple generations
Chip Development Process

Business Case
- Marketing

Program Commit
- Engineering

RTL
- Design

Verification
- Arch
- block

Netlist
- Timing Closure

Layout
- Floorplan
- Timing

Tapeout
- checklists

Samples
- System Verification

Production Release
- Volume Production

© 2019 Broadcom. All Rights Reserved. The term “Broadcom” refers to Broadcom Inc.
TOMAHAWK FAMILY
Three High Performance Switch Architectures - Broadcom

- 10.0T: Massive Bandwidth for Hyperscale Fabrics
- 5.0T: Programmable, Feature-Rich Switches for Enterprise and Data Center
- 1.0T: Scale-Out, Converged Carrier-Grade Infrastructure

Bandwidth Features

- Tomahawk
- Trident
- Jericho

© 2019 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc.
Scaling up the Network with Merchant Silicon

- **2010**: 64x serdes 10G NRZ, 0.64Tbps
- **2012**: 128x serdes 25G NRZ, 1.28Tbps
- **2014**: 128x serdes 25G NRZ, 3.2Tbps
- **2016**: 256x serdes 25G NRZ, 6.4Tbps
- **2018**: 256x serdes 50G PAM-4, 12.8Tbps
- **2020**: 50G or 100G PAM-4, 25.6Tbps

**40X Bandwidth Increase per Switching Element Over 10 Years, Exceeding Moore’s Law**

© 2019 Broadcom. All Rights Reserved. The term “Broadcom” refers to Broadcom Inc.
Data Center Market

Accelerating 25/100GbE in the Data Center

- 100G has a Long Tail
- 25G will replace 10G in Server Access
- 40G continues to decline

Source: Dell'Oro Oct 2017 Tables
650 Group 2017 Report

Source: https://www.max.ee/sites/default/files/dell_open_networkingVision_and_portfolio.pdf
Tomahawk 3: By the numbers

- 12.8 Tb/s multilayer Layer3 switching
- Configurable as 32x 400GbE, 64 x 200GbE, or 128 x 100GbE
- 256 dual-mode – 56G-PAM4 and 28G-NRZ
- 40% Power reduction per 100GbE port
- 75% lower cost per 100GbE port
- Integrated shared-buffer architecture
- Broadview Gen3 network instrumentation
- IP forwarding, ECMP
- Dynamic Load Balancing and Group Multipathing
- In-band Network Telemetry
- 16 nm process geometry
- In Production now

Source: https://www.broadcom.com/blog/broadcom-s-tomahawk-3-ethernet-switch-chip-delivers-12-8-tbpd-of-speed-in-a-single-16-nm-device
Tomahawk 3 Architecture

Source: https://www.linleygroup.com/mpr/article.php?id=11908

© 2019 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc.
Terminology

• VLAN – Virtual LAN
  • Virtual LAN

• L2 Table
  • Table looked up with key = Destination MAC address
  • Determine the outgoing port

• L3 Table
  • Table looked up with key = Destination IP address
  • Determine the outgoing interface/port

• ACL – Access Control List
  • Implements access control policies
Packet Receive → Check framing, CRC → Packet header fields are parsed → VLAN assignment → Destination MAC Address Lookup \rightarrow Port → Destination IP Address Lookup \rightarrow Port → Access Control Lists → Packet queued to output port

MAC → Parser → VLAN Table → L2 Table → L3 Table → ACL Table

Packet Receive → Packet receive → Check framing, CRC → Packet header fields are parsed → VLAN assignment → Destination MAC Address Lookup \rightarrow Port → Destination IP Address Lookup \rightarrow Port → Access Control Lists → Packet queued to output port

Packet Buffer → ... → Packet receive → Check framing, CRC → Packet header fields are parsed → VLAN assignment → Destination MAC Address Lookup \rightarrow Port → Destination IP Address Lookup \rightarrow Port → Access Control Lists → Packet queued to output port

Packet Edits → Egress ACL → MAC

© 2019 Broadcom. All Rights Reserved. The term “Broadcom” refers to Broadcom Inc.
Example: ECMP Load Balancing

Packets steered based on Flow Hashing
- Distribute flows equally among links as much as possible
- Switch chip should have capability to provide
  - Sufficient depth of parsing
  - Hashing
  - Ability to handle different types of flows
Tomahawk 3 enables Cost and Power Reduction

75% Reduction in System Power, 85% reduction in System Cost *

Source: Facebook, OCP

Power Metric Includes Optics, Cost Metric Excludes Optics
Industry’s Broadest Ecosystem

**Network Controller**
- OPEN DAYLIGHT
- lago
dus
- nuage networks
- Ryu
- NEC
- Juniper
- OPEN CONTRAIL

**NV Controller**
- PLUMgrid
- midokure
- VMware NSX

**Orchestration & Automation**
- openstack
- CHEF
- puppet labs

**Operating System**
- Arista
- HP
- Cisco
- Juniper Networks
- Extreme networks
- H3C
- Huawei
- ZTE

**Hardware Platform**
- Celestica
- Open Company
- DNI
- Accellena
- Arista
- ALPHA Networks
- HPE
- H3C
- Huawei
- ZTE

© 2019 Broadcom. All Rights Reserved. The term “Broadcom” refers to Broadcom Inc.
Key Takeaways

• Switch Silicon development is about 18 to 24 month process

• Requires investment of 50 – 100 million dollars

• Cooling techniques are challenging and expensive

• Process Geometry is not yielding cost and power advantage

• Monolithic dies may be replaced with multi-die in a package
thank you