



PUSHING THE LIMITS, A PERSPECTIVE ON ROUTER ARCHITECTURE CHALLENGES

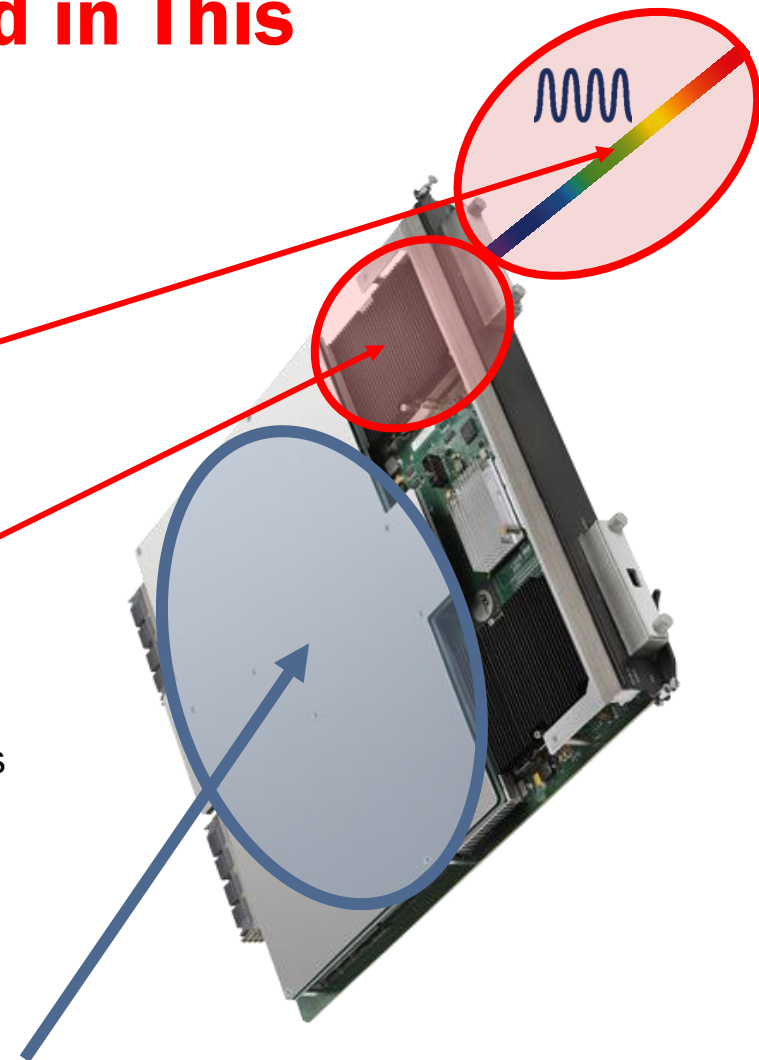


Greg Hankins <ghankins@brocade.com>

NANOG53

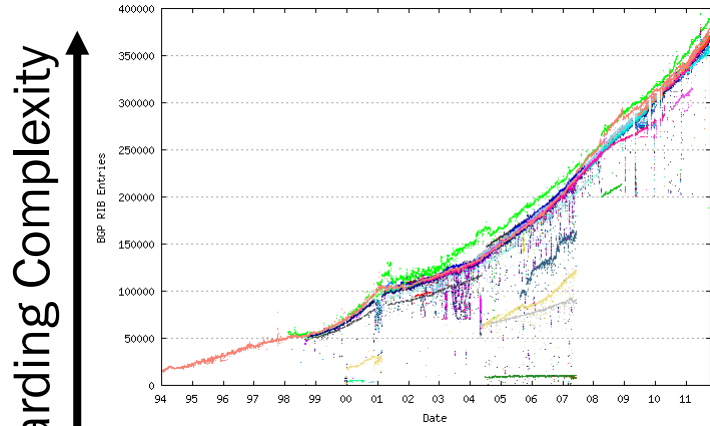
Agenda and What's Covered in This Presentation

- At the June NANOG 52 in Denver we covered
 - Optical technology: “Dawn of the Terabit Age: Scaling Optical Capacity to Meet Internet Demand”
<http://www.nanog.org/meetings/nanog52/abstracts.php?pt=MTgwOSZuYW5vZzUy&nm=nanog52>
 - Ethernet interface technology: “100 GbE and Beyond”
<http://www.nanog.org/meetings/nanog52/abstracts.php?pt=MTc2MSZuYW5vZzUy&nm=nanog52>
- We haven't talked about router hardware architectures in detail since NANOG 39 in Toronto (2007)
 - BoF: Pushing the FIB limits, perspectives on pressures confronting modern routers
<http://www.nanog.org/meetings/nanog39/abstracts.php?pt=MjY4Jm5hbm9nMzk=&nm=nanog39>
- Focus of this presentation is on memory and ASIC technology that is used to continue to meet the increasing router packet processing, lookup capabilities and memory scalability requirements

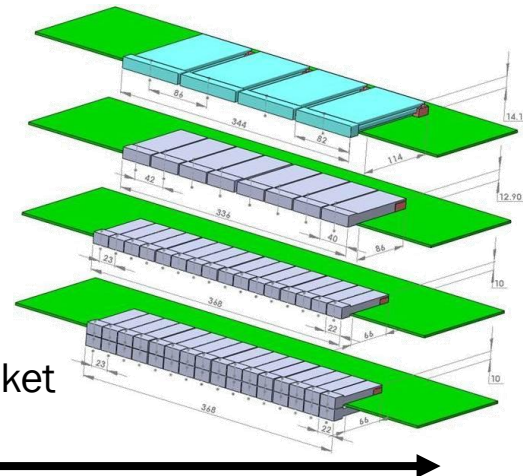
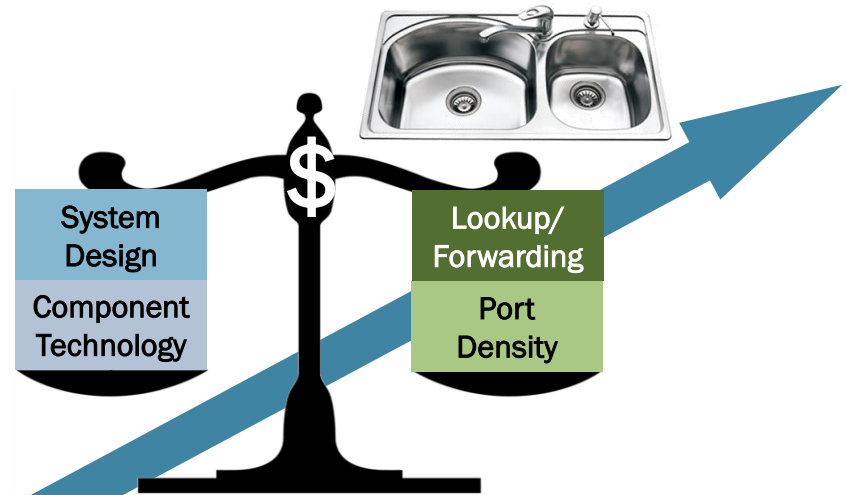


Lookup and Forwarding Hardware Design Challenges

Solutions are Good, Fast, or Cheap – Pick any Two



- Complex multi-protocol forwarding
- Growing IPv4/IPv6 Internet and VPN routing tables
- IPv4 deaggregation and IPv6 adoption make future growth hard to predict

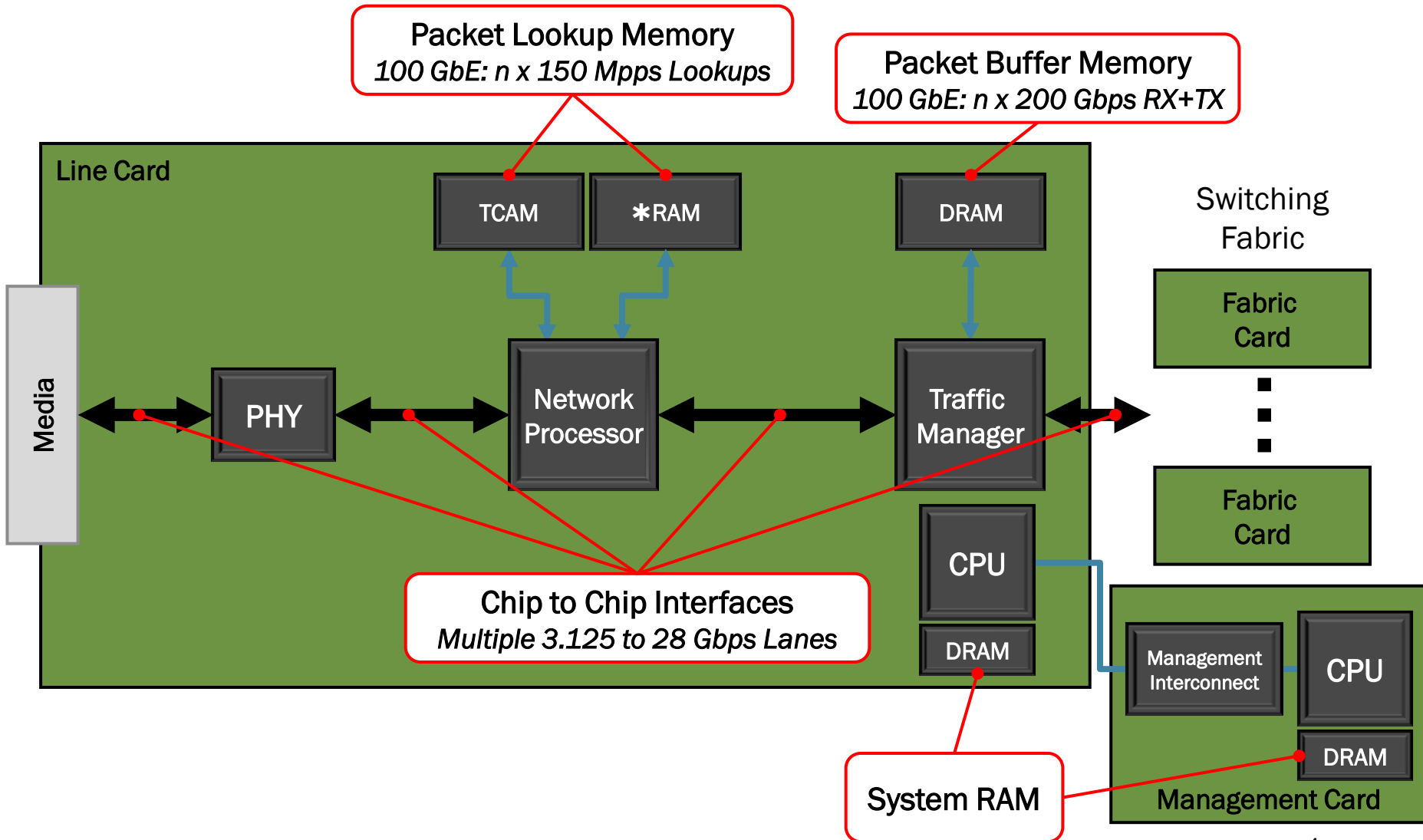


- High density 100 GbE
- Line rate 100 GbE is one packet every 6.72 ns

Lookup Capacity and Forwarding Complexity

Port Density

Basic Router Forwarding Architecture



Agenda

- Introduction
- **Memory Technology**
- ASIC Technology
- Summary

Key Memory Challenges

- Packet rates have greatly exceeded memory random read rates
 - Lookup: 1 ns random read rates needed yesterday
 - Buffer: 1 ns random read and writes rates needed yesterday
- Dynamic memory technology characteristics impose significant constraints on lookup and buffering architectures
 - Inherent restrictions and non-random access
 - Bank blocking due to previous read/write events
 - Applies to both on-chip and off-chip solutions
- Today's solutions offer
 - ~48 ns random read rates for commodity off-chip DRAM
 - ~20 ns random read rates for specialized off-chip RLDRAM
 - ~2 ns random read rates for on-chip embedded memory

Lookup and Buffering Memory Requirements



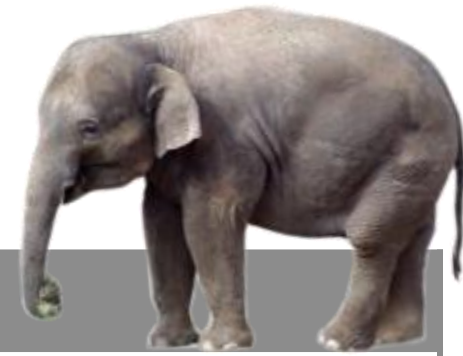
Fast!

- Have to store everything needed for packet lookups in hardware to forward at line rate with all features enabled
 - 10 GbE: 15 Mpps or one packet every 67 ns
 - 100 GbE: 150 Mpps or one packet every 6.72 ns
 - Multiple ports on a network processor
 - Multiple lookups are needed per packet



Big!

- Packet lookup tables hold
 - MAC address table (L2, VPN)
 - IPv4 FIB (unicast and multicast, VPN)
 - IPv6 FIB (unicast and multicast, VPN)
 - VLAN tags, MPLS labels
 - ACLs (L2, IPv4, IPv6, ingress and egress)
 - QoS policies (PHB, rewrite, rate limiting/shaping)
- Deep buffering, queuing and shaping
 - Multiple 100 Gbps of sustained throughput into buffer memory
 - 1 GB buffer is only 80 ms at 100 GbE rates
 - 1000s of queues per port

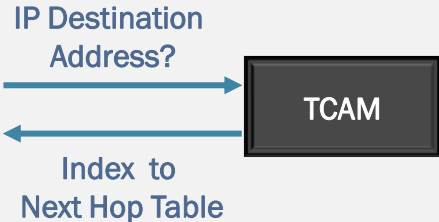
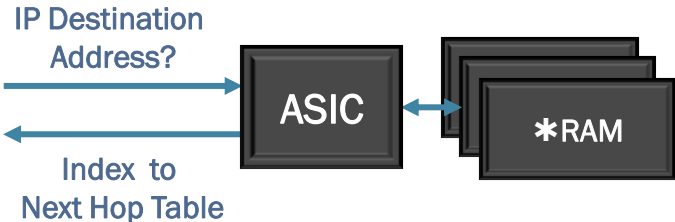


I'm Fast and Big,
But I Don't Exist

Lookup and Buffering Memory Technology Overview

	TCAM	SRAM	DRAM		
			DDR	RLDRAM	GDDR
What is it?	Ternary Content Addressable Memory	Static RAM	Double Data Rate Dynamic RAM	Reduced-latency Dynamic RAM	Graphics Double Data Rate Dynamic RAM
Primary Function	Wildcard Search	High Speed Storage	High Capacity Storage	High Speed Storage	High Bandwidth Storage
Industry Usage	Specialty	Commodity	Commodity	Specialty	Specialty
Access Speed	Lower	Lowest	Highest	Higher	Higher
Density	Lower	Lower	Much Higher	Higher	Much Higher
Cost per Bit	Much Higher	High	Much Lower	Lower	Lower
Power Consumption	Highest	Lower	Higher	Higher	Higher
Mass Production Capacities	40 Mbit Today, 80 Mbit Soon	72 Mbit Today, 144 Mbit Soon	2 Gbit DDR3 Today, DDR4 Soon	576 Mbit Today, 1 Gbit Soon	2 Gbit GDDR5 Today

Longest Prefix Matching (LPM) Mechanisms

	TCAM	*RAM
What is it?	 <p>Wildcard Hardware Data Search</p>	 <p>Ordered Tree Data Structure Search in Hardware</p>
Cost	Much Higher	Lower
Power Consumption	Higher	Lower
Search Latency	Fixed	Variable
Throughput	Higher	Lower, Must Parallelize
Add/Delete Time	Fixed	Variable
Prefix Capacity	Fixed, Lower	Variable, Higher
Search Algorithm	Ordering of Data in TCAM	Lots of Different (Patented) Algorithms with Tradeoffs
Software Complexity	Versatile Data Storage No Need to Worry About How You Write the Data	Need to Implement Lookup and Data Structure Mechanisms Separately Must Optimize Layout for Data Structure

LPM Memory Architecture Solutions

- Divide and conquer parallel architecture
 - Deterministic search using a combination of SRAM/DRAM
 - Large number of banks allows parallel searching in reasonable time
- Integrate lookup memory into packet processing ASICs
 - Combine embedded memories for higher performance and reasonable density
- Component technology must be available for 5+ years at a minimum
 - Many specialized memory technologies have a window of production that is too small

Buffering Memory Solutions

- DRAM read and write times are the limiting factor in guaranteeing buffering performance
 - Commodity off-chip DDR4 DRAM will give us higher memory throughput
 - Embedded on-chip memory limits buffer sizes but offers higher performance
 - Proprietary buffer memory management techniques
- Economically and technically feasible to design a custom buffer memory chip, instead of using commodity DRAM

Agenda

- Introduction
- Memory Technology
- **ASIC Technology**

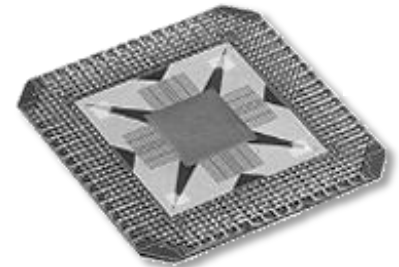
Key ASIC Challenges

- Multi-protocol packet forwarding requires complex packet parsing and deep header inspection
 - L2, IPv4/IPv6, unicast/multicast, broadcast, 802.1Q, 802.1ah, TRILL, 802.1aq, MPLS
 - LAG load balancing requires complex hashing functions and looking deep into multiple packet headers
 - sFlow, OAM (hardware timestamps)
 - Tons of counters (virtual interfaces, aggregate VLAN)
- It's all about integration – the more functionality you can put on one chip, the more scalable you can make the system
- Available process geometry, packaging and chip interconnects limit what we can do today
 - Using more chips in parallel means higher cost, higher power consumption, higher heat dissipation and a lower MTBF
 - Current generation ASICs use 45 nm and 32 nm technology
 - Industry is moving to 22 nm and smaller as soon as it's technically and commercially feasible
 - Current SERDES and chip interconnects are limited to ~6.25 Gbps, 10 Gbps between chip and optics, 28 Gbps available soon
 - Our 2-port 100 GbE card uses around 100 ICs, need to reduce the number of components!

Semiconductor Processes Geometries

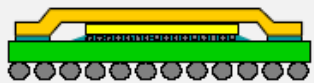
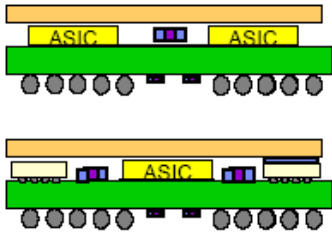
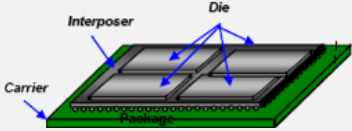
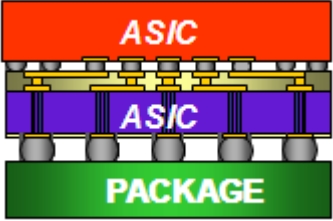
Moore's Law: The Number of Transistors on a Chip Will Double Approximately Every Two Years

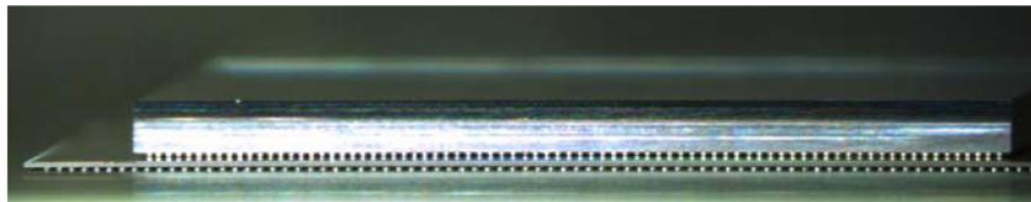
Process Geometry	180 nm	130 nm	90 nm	65 nm	45 nm	32 nm	22 nm	16 nm	11 nm
Availability	1999	2000	2002	2006	2008	2010	2011	~2013	~2015
Transistors (Billions)	0.06	0.18	0.52	0.73	1.77	4.8			
Gates (Millions)	2	6	15	21	30	50			
Frequency (MHz)	106	212	360	400	450	750			
Memory (MB)	6	49	164	72	315	900			
					In Use Today				



- Process geometry refers to the smallest dimension that can be drawn into the silicon to define a transistor
- Transistors are the building blocks used to make gates
- Maximum capacities are often proprietary to a vendor, examples are from actual Brocade ASICs

ASIC Technology Evolution

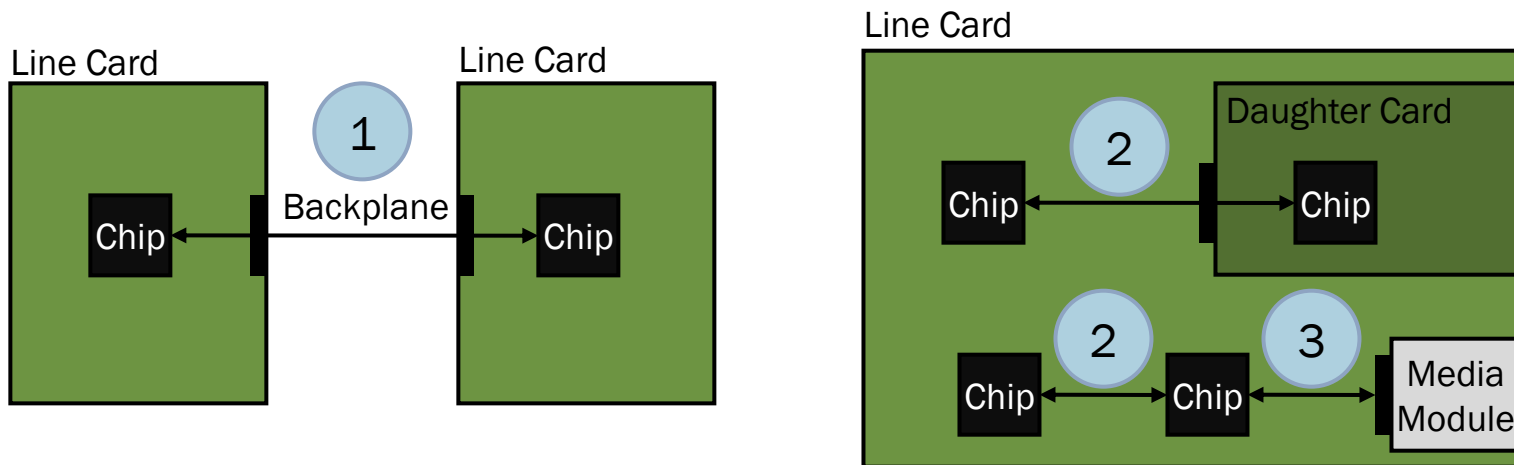
	2D Single Chip Package	2D Multi-chip Package	3D Silicon Interposer	3D Package
What is it?				
	Single Large Chip Package	Multiple Chips or Components in the Same Package	Multiple Chips in the Same Package With a Silicon Interconnect	Two Chips in a Stack in the Same Package
Size	Larger	Larger	Smaller	Smaller
Interconnect	Longest , External Off-chip	Shorter, Copper/PCB	Shorter, Silicon	Shortest, Vertical
Power	Highest	Lower	Lower	Lowest
Heat Density	Lowest	Higher	Higher	Highest
Signal Density	Lowest	Higher	Higher	Highest



IBM 3D Package

Chip Interconnect Signaling

- Evolutionary step in SERDES technology standardized by the OIF
 - 25/28 Gbps electrical signaling will make newer chip to chip interfaces and pluggable media modules possible
1. Backplane: CEI-25G-LR – 30”
 2. Chip to chip: CEI-28G-SR – 12”
 3. Chip to module: CEI-28G-VSR – 4”
(used by 2nd generation 100 GbE media modules)



ASIC Technology Integration

- Current process technology is optimized for either logic or memory requirements
 - ASIC technology is logic driven (network processors)
 - Memory technology is memory driven (RAM)
 - 3D multi-chip packaging allows us to do both in the same package with optimized chips for each function
- When a new process geometry is developed, component vendors focus on transistor density first
 - The rest of the technologies (SERDES, on-die memories, etc.) in the logic library are developed afterwards
 - Typically used for general-purpose CPU production first
 - Even though a process geometry may be commercially available, we have to be able to integrate it

ASIC Technology Integration

- Board layout architecture must be carefully planned together with memory and ASIC components
 - Board real estate
 - Power consumption
 - Heat dissipation
 - Signal integrity and routing between components
- Advanced 2D and new 3D ASIC technology will simplify the board layout challenges
 - Expected to be mainstream within the next few years
 - Combine multiple ASICs or memories within the same package
 - Shorter interconnect and lower latency between components (especially memory)
 - Lower board power consumption and heat dissipation
 - Uses less board real estate



The Integration Challenge
*Main Board and Daughter Board
Component Density vs. Real Estate*

Agenda

- Introduction
- Memory Technology
- ASIC Technology
- **Summary**

Alternatives to FIB Scaling

- Challenge is using available technology vs router density and complexity to continue to scale

- Moore's Law still holds true

- Software FIB aggregation or suppression mechanisms

- “On the Aggregatability of Router Forwarding Tables”

- <http://www.cs.arizona.edu/~bzhang/paper/10-infocom-aggregate.pdf>

- Simple Virtual Aggregation (S-VA)

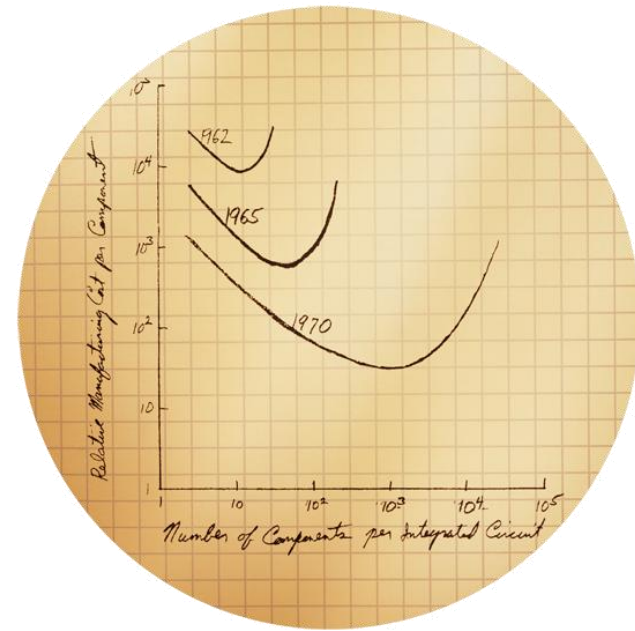
- <https://datatracker.ietf.org/doc/draft-ietf-grow-simple-va/>

- OpenFlow/Software-Defined Networking (SDN)

- <https://www.opennetworking.org/>

- Locator/ID Separation Protocol (LISP)

- <https://datatracker.ietf.org/wg/lisp/charter/>



Summary

- Want to design hardware with a minimum of 7+ years of lifetime that supports
 - Multi-protocol forwarding
 - High density 100 GbE with CFP2 optics
 - Scalable and flexible FIB (global routing table + VPN)
- Requirements dictate solution complexity and cost
 - Technology capabilities
 - Functionality and port density
 - Design tradeoffs have to be chosen carefully
- Higher capacity and faster components are coming that will allow us to continue to scale router density, lookup capacity and memory requirements
 - Multi-100 Gbps packet processing ASICs
 - Custom SRAM/DRAM lookup memory
 - Custom DRAM packet buffering memory
 - 3D technology with ASIC and memory chips in the same package



Questions?

