Memories for Extreme Networking
Memcon 2014

Michael Sporer
Director of Marketing
msporer@mosys.com
Topic Suggestions

- Latency challenges and solutions?
- Bandwidth versus capacity challenges?
- On chip versus off chip?
- What does the memory system look like …
  - size, throughput, power, energy
Agenda

- Background

- Packet Buffering and Header Processing
  - go hand in hand → but they are not the same

- Next Generation for Extreme Networking

- Memories for Header Processing
Available, Multi-sourced commodity memories

- **DRAM**
  - Planar Cell → Stack cell vs. Trench cell vs. Crown

- **NVM**
  - MROM/OTP/EEPROM → Flash (NOR vs. Mirror-bit vs. NAND)

- **SRAM**
  - ‘Slow’ SRAM for low density and low power applications. Simpler and cheaper than DRAM.
  - ‘Fast’ SRAM for caches
  - In the early 90’s embedded was expected to wipe out SRAM …

Networking specific memories did not exist in the early 1990’s
Memory Core Design

- **DRAM Array Performance**
  - In 1994 tRC was ~110ns, except for IBM mainframe DRAM (<50ns)
    - IBM was a captive producer of DRAM for their internal consumption

- **Multi-bank DRAM architecture emerged with SDRAM**
  - tFAW which reduces random access performance came later

- **Soft error susceptibility of DRAM and SRAM cells**

- **Row Hammer**
  - Row hammer sensitivities were discovered in the earliest DRAMs and were designed out.
  - Introduction of CPU caching hierarchy eliminates Row Hammer type access in Compute applications.
Memory Interface Performance

- **DRAM**: FPM → EDO → SDR → DDR – improved bandwidth
  - Dual port VRAM/WRAM (RA + SA) for frame-buffers
  - Early SDRAM was used in workstation texture mapping (16Mb)

- **Many DRAM innovations were proposed; some produced**
  - Burst EDO
  - Virtual Channel
  - ES-DRAM / C-DRAM
  - 3D-RAM
  - RDRAM
  - SLDRAM
  - DRDRAM
  - FCRAM – for networking – 25ns tRC

- **SRAM**
  - From Asynchronous >10ns → to Synchronous >200 MHz
Commoditization drives volume up and cost down
- The cost advantage becomes so great that applications are optimized for the memory.

Extreme Networking – Specialty Memory
- Memory is optimized for the application
- Delivers a compelling performance advantage.

Networking memories emerged in the 1990’s and have evolved:
- SRAM
  - Async → Flow-through → ZBT et.al. → QDR & Sigma Quad
- DRAM
  - FCRAM (fast cycle RAM)
  - RLDRAM (reduced latency DRAM)
  - LLDRAM (low latency DRAM)
- Content Addressable Memories (CAM, TCAM):
  - Architected for searching a table via a match comparator
- Embedded (On-chip) Memories
Agenda

- Background

- Packet Buffering and Header Processing
  - go hand in hand → but they are not the same

- Next Generation for Extreme Networking

- Memories for Header Processing
Packet Processing

Packet Header Processing

- Large packets (ie. streaming) have low packet rate and approach the datarate of the port
  - Defines maximum Buffer date rate

- Small packets (ie. acknowledge) potentially have a high packet rate but no data throughput
  - Defines maximum header processing rate

- Must design to support both
  - Maximum data rate
  - Maximum packet rate
Packet Processing Rates

- **Looking ahead to 400G ports**
  - Longer header of IPv6 reduces the maximum packet arrival rate
    - Packet rate still increasing, but not as quickly
L2/L3 Packet Header Processing

- Lots of memory transactions and many different tables
- Packet buffer not illustrated

**Figure 12:** Packet forwarding pipeline stages inside a Arista 7250X / 7300X

Source: Arista
Broadcom Trident II – 1.28 Tbps

- 32x40G / 96x10G + 8x40G
- 388Mb memory
  - 96Mb packet buffer (dynamic)
  - 292Mb for table store, etc.

Top of Rack / Aggregation – 10G w/ 40G uplink

Modular Chassis – all 40G
  - Forwarding and Fabric

Source: Broadcom
Example Leaf and Spine Network

- These switch chips used for forwarding and fabric within Leaf and Spine Architecture

- Cisco Datacenter Example:
  - Cisco: Hybrid = merchant plus custom silicon
  - Configuration Options:
    - Layer 3 IP fabric
    - Layer 3 all the way to access
  - All memory is on-chip

Source: Cisco live!
It all boils down to this

- Single Chip Solution
- Highest Throughput ...
- Lowest Latency ...
- Large Leaf and Spine networks ...
- So, what’s missing?
... quite a bit actually

- Classification
- Traffic Management
- Service Level requirements
- ... and the list goes on

- Whatever the workload, more sophisticated PP requires
  - more memory transactions per packet
  - more tables and/or larger tables
  - and larger buffer time

- On-chip memory cannot meet these capacities

<table>
<thead>
<tr>
<th></th>
<th>External DRAM</th>
<th>On-chip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Arista 7508E</td>
<td>Arista 7050</td>
</tr>
<tr>
<td>Maximum Scale</td>
<td>55,0000+ compute nodes</td>
<td>2,000 compute nodes</td>
</tr>
<tr>
<td>Per port buffering</td>
<td>128MB (100ms)</td>
<td>5MB (dynamic)</td>
</tr>
</tbody>
</table>
Advanced Packet Processing

Packet Header Processing

- Classification / Lookup
- Flow State Retrieval / Update
- Statistics
- Metering / Policing
- Congestion Control
- Per Flow Queuing
- Shaping

‘Deep’ Packet Buffer
Tale of Two Switches

- **Core Network Switch Line Card**
  - External DRAM Packet Buffer (2x16)
  - External TCAM (4x2)
  - 480 Gbps throughput

- **Datacenter Leaf Switch**
  - All on-chip memory
  - 960 Gbps throughput
Broadcom 88650
- 200 Gbps Line Rate
  - 240 Gbps oversubscribed
  - 24x10G transceivers
- 16 DDR3 DRAM @ 2133 Mbps
  - 37% over-provision BW
  - 133 ms buffer capacity
- Option for external lookup
- Option for external statistics

Assume:
- Body Size: 45x45 mm

Next Generation – probably 400G, maybe 500G

Next+1 Gen → assume 800G datapath
- Also assume: 100ms* buffer & 37% BW overprovision requirement

* - Historical Precedence
Agenda

- Background

- Packet Buffering and Header Processing
  - go hand in hand → but they are not the same

- Next Generation for Extreme Networking

- Memories for Header Processing
**DDR3 → DDR4**
- 2133 → 3200
- Bandwidth Constrained
- 42 DRAM
- Body Size: 55x55 mm
  - ~2900 balls

- Only half the DRAM shown

**Not feasible**
- Not routable
### 800G Packet Buffering with 3D TSV DRAM

<table>
<thead>
<tr>
<th>Attribute:</th>
<th>HBM</th>
<th>HMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM Layer Density</td>
<td>8Gb?</td>
<td></td>
</tr>
<tr>
<td>Stack Height</td>
<td>4H or 8H</td>
<td></td>
</tr>
<tr>
<td>tRC (cycle time)</td>
<td>same (DRAM-ish)</td>
<td></td>
</tr>
<tr>
<td>IO (rate)</td>
<td>Parallel (1-2G)</td>
<td>Serial (up to 25G)</td>
</tr>
<tr>
<td>Bandwidth Granularity</td>
<td>8 channel</td>
<td>One Link = 16 lane</td>
</tr>
<tr>
<td>Package Granularity</td>
<td>8 channel</td>
<td>32 or 64 lane</td>
</tr>
<tr>
<td>Package Mounting Type</td>
<td>Si-interposer</td>
<td>Discrete</td>
</tr>
<tr>
<td>100 ms Packet Buffer Capacity</td>
<td>4H = 300 Gbps, 8H = 600 Gbps</td>
<td>Up to 640 G per 32 lanes</td>
</tr>
<tr>
<td>Throughput per stack</td>
<td>365~730 Gbps</td>
<td>Up to 640 G per 32 lanes</td>
</tr>
</tbody>
</table>

- **Neither solution is optimum for 800 Gbps**
  - Applications are optimized for the memory

- **600 Gbps is stepping backwards so → 1200 Gbps**
Datapath = 1200G = 48 x 25G

HMC @ 80% efficient 25G → 64 transceivers
- 112 total transceivers
- Requires Body Size: 50x50
- ~300 extra pins available

100ms = 2x64Gb = 2x8H stack
- 2x cubes @ 32 lanes each
- Assume 16x19 (0.65 mm pitch) HMC package
  - Serial links ease routing
  - Reduces thermal density
Datapath: 1200G = 48 x 25G

HBM @ 1.7Gbps/pin
- Includes 37% overprovision
- Headroom to 2Tbps
  → supports Oversubscription

Assume maximum interposer size
- Body Size: 45x45
- ~400 pins available on the package

100ms = 2x64Gb = 2x8H stack
Deep (100ms) Packet Buffer Solutions

PB defined by throughput
- Reducing the header processing complexity does not reduce the PB

Header processing can have a wide range of configurations
- Depends on the needs of the applications

Power & Thermal Density
- Large Packet Processors consume substantial power
- How to Reduce the power without reducing throughput?
  → Solution: VSR links to external PHY chips
    • MLG Gearbox for legacy 10G interconnect,
    • Retimer for backplane, optics modules

<table>
<thead>
<tr>
<th>Memory</th>
<th>Board Area</th>
<th>extra pins</th>
<th>IPv6 Packet Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDR4 (800G)</td>
<td>~7400</td>
<td>~0</td>
<td>&lt;600M/s</td>
</tr>
<tr>
<td>HMC (1200G)</td>
<td>~3100</td>
<td>~300</td>
<td>&lt;900M/s</td>
</tr>
<tr>
<td>HBM (1200G)</td>
<td>~2000</td>
<td>~400</td>
<td>&lt;900M/s</td>
</tr>
</tbody>
</table>
Agenda

- Background
- Packet Buffering and Header Processing
  - go hand in hand → *but they are not the same*
- Next Generation for Extreme Networking
- Memories for Header Processing
### Memory Transactions per Packet

Will use both on and off chip: 50:50
- 900 Mpps x 12 → ~11 Btps on chip
- 900 Mpps x 12 → ~11 Btps off chip

**Searches:**
- 900 Mpps x ~3 → ~3 Bsps

<table>
<thead>
<tr>
<th>IP processing features</th>
<th>IPv4 Core QoS</th>
<th>IPv6 Core QoS w/o policing</th>
<th>IPv6 Core QoS w/ policing</th>
<th>IPv4 Edge RPF</th>
<th>IPv4 Peering</th>
<th>MPLS Switch</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIDB/Index table RAM read</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Dest/Src Lookup (prefix+table) RAM read</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>Stat RAM RMW</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Policing RAM RMW</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RED RAM read</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RED max enQ/deQ RAM RMW</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>ACL range lookup</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td><strong>TCAM search or lookup (ACL, QoS, ...)</strong></td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>TCAM associated RAM read</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total RAM Transactions</strong></td>
<td><strong>20</strong></td>
<td><strong>19</strong></td>
<td><strong>26</strong></td>
<td><strong>25</strong></td>
<td><strong>24</strong></td>
<td><strong>18</strong></td>
</tr>
</tbody>
</table>
A 1200G HMC based Packet Processor

- **900 Mpps (tRC = 1.1ns)**
  - Can support small tables with on-chip memory

- **On Package Memory?**
  - Each config is a custom package and memory?
    - Si Interposer required?
  - High transaction memory is high power

- **Off Package ~300 Unused IO pins**
  - More Datapath?
    - Oversubscription – need more buffer BW
      - Add 16 lane HMC for OS supports 12 lanes DP
  - Off-Package Memories (900Mpps)
    - Parallel – SRAM / RL / LLDRAM – not enough pins
    - Serial TCAM – limited range of functionality and capacity
    - Bandwidth Engine – unified memory solution:
      - Search, Lookup, Metering, Statistics
    - Off package memories compete with OS datapath and buffer
    - Adds board area
    - More pins = larger host package
900 Mpps (tRC=1.1ns)
- Can support small tables with on-chip memory

~400 Unused IO pins
- More Datapath
  - Oversubscription
  - >20% possible with HBM speed-up
- Parallel Memories
  - SigmaQuad SRAM – delivers highest random memory transaction rate and lowest latency
- Serial Memories
  - Serial TCAM – limited range of functionality and capacity

Transceivers are a fungible resource, parallel pins are not
1.2 Tbps Line Rate, Multi-Config Switch

- Oversubscription (datapath)
- Additional serial interface devices

Datapath ICs

Optics Modules

GB/RT

MoSys LineSpeed Gearbox Retimer

HBM

SigmaQuad-IV

SRAM

Bandwidth Engine-3

45x45

MoSys

LineSpeed Gearbox Retimer

Ultimate Header Processing Memory Solutions

MoSys

Bandwidth Engine-3

MoSys

SigmaQuad-IV SRAM

MoSys

Bandwidth Engine-3

MoSys

LineSpeed Gearbox Retimer

SigmaQuad is a trademark of GSI Technology
Bandwidth Engine and Sigma-Quad SRAM

**BE-2: In Production**

- **Parallel Internal Architecture**
  - Concurrent Read/Write

- **Network Offload Acceleration**
  - Single Chip Buffering
  - Statistics, Metering, Atomic Operations, Search

- **Ultimate End to End Data Protection**
  - ECC, CRC, Bit Safe Self Test and Self Repair
  - On-chip ECC (Sigma-Quad)

- **Second Source Available Now**
  - GSI Technology

---

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Bandwidth Engine</th>
<th>Sigma-Quad IVe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BE-2</td>
<td>BE-3</td>
</tr>
<tr>
<td>Transaction Rate</td>
<td>4.5 B</td>
<td>&gt;10 B</td>
</tr>
<tr>
<td>Throughput (full duplex)</td>
<td>200Gbps</td>
<td>400Gbps</td>
</tr>
<tr>
<td>Capacity</td>
<td>576 Mb</td>
<td>1152 Mb</td>
</tr>
<tr>
<td>Random Cycle Time</td>
<td>2667 ps</td>
<td>2667 ps</td>
</tr>
<tr>
<td>Interface</td>
<td>GCI</td>
<td>GCI</td>
</tr>
<tr>
<td>Intelligent Offload Macros</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Deterministic Random Access</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multi-Cycle Macros</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
The Bandwidth Engine delivers the Highest Performance of any External Memory

- Most Efficient GCI Protocol
- 25G+ IO Rates x 16 lanes
- Up to 32 concurrent memory ops
- Highest Look-Up Performance ➔ up to 6 billion reads per second
- Multi-function, Unified Memory

Single Cycle Macros (Atomic RMW)

Single Chip Buffer (400G port)

Multi-Cycle Macros (Algorithmic)

The Bandwidth Engine delivers the Highest Performance of any External Memory

- Most Efficient GCI Protocol
- 25G+ IO Rates x 16 lanes
- Up to 32 concurrent memory ops
- Highest Look-Up Performance ➔ up to 6 billion reads per second
- Multi-function, Unified Memory
Intelligent Memory Macros

- Offload the host of memory intensive operations
  - One macro instruction replacing 2 to >20 individual accesses
  - Full memory bandwidth & transaction rate with 16 lane interface
  - Macros utilize multi-cycle and parallel pipeline capabilities with 8 or 4 lanes
  - Macros execute Atomically
    - Statistics, Metering, Read & Set, Test & Set, Algorithmic Operations

Minimize Data Movement = Increased Efficiency
Bandwidth Engine 3
Also scales down to 400G and 200G
MoSys Delivers:

- **Bandwidth Engine Product Family**
  - Offload Acceleration → Unique in the Industry
    - Removes bottlenecks by minimizing data movement to and from the host
    - Single Cycle and Algorithmic Macro Functions
  - Highest Transaction Rate and Data Throughput in One Chip

- **LineSpeed Data Path ICs**
  - Range of Retimer, Gearbox and other PHY based solutions
  - Highest Performance per Watt in the Industry

- **Carrier Grade Quality and Reliability**
  - Designed and manufactured with customer requirements in mind
  - Bandwidth Engine - Bit Safe™ Self Test and Self Repair Option

- **System Value → Replaces multiple legacy memory devices**
  - Delivers tangible Power, Footprint and Economic advantages
  - Efficient Serial Interface and GCI Transport Protocol
Latency challenges and solutions?
- It’s no longer about latency, transaction rate is the new key metric

Bandwidth versus capacity challenges?
- Advanced packaging such as HBM or HMC enables DRAM to scale beyond DDR4

On chip versus off chip?
- Each is needed depending on the application

What does the memory system look like …
- size, throughput, power, energy?
- Technology may enable the size, throughput, power, energy

→ Engineering determines configurability, efficiency and cost effectiveness
Thank You
Backup

Memcon 2014
Effective Memory Size vs Speed Up

- Using multiple copies of tables can mimic a faster memory device, but with diminishing returns.
Full Duplex Networking Architecture

- Effective bandwidth of the buffer is 2x the line rate
  - Conservation of data (Out = In) = 2x Line Rate
  - Memory Bandwidth = (In + Out) / %efficiency

- Header processing must occur at least as fast as the packet arrival rate and the buffer capacity
  - Arrival Rate = Line Rate (bits per time) / Packet Size (bits per packet)
  - Buffer Capacity (time) = Buffer Size (bits) / Line Rate (bits per time)
Packaging - Advancing

- Embedded Interposer
  - Recently disclosed by Intel foundry
  - Eliminates TSVs
  - Introduces other challenges
  - Long term: will reduce cost

Silicon – Slowing

- extended product lifecycles
  - The same will be true for DRAM
  - Lower risk to designing in an MCM solution like HBM.
32 port x 100 Gbps switch chips
- 25G transceivers
- 3.2 Tbps throughput
- All on-chip memory
  - Limited time for decision
- Low Latency