Billions of Packets per Second

David Mendel
Altera Corporation
Billions of packets per second...

- 1Tbps = 1.6 Billion packets/second
  - 40Gbps/100Gbps commonly requested today
  - 400Gbps active research with 1Tbps roadmap

- Complexity continues to grow
  - “The nice thing about standards is that there are so many to choose from”
  - Continuing new and updated standards, while continued backward compatibility required; strong value in early adaption

- Customers want it their way
  - Desire to differentiate requires customization and works against ASSPs
Challenges

**Speed & Throughput**
- Ten years ago, 1.25Gbps serdes introduced in an FPGA
  - 11.3Gbps shipping in volume today
  - 28Gbps coming; test chips demonstrated
  - More to come…
- Allows PP/TM to move from 40Gbps/100Gbps of today to 400Gbps/1Tbps of tomorrow

**Power / thermal**
- Think inside the box
- Next gen line-cards will go in existing box; limited power and cooling
- Need more throughput at similar power
Challenges

- **Memory**
  - Limit to how many pins you can put on a package
  - 400Gbps needs > 1.8Tbps memory throughput for packet buffering
    - ~2700 pins + PWR/GND
  - 1Tbps needs > 5000 I/O pins

- **What to harden?**
  - Hardened logic is more efficient (power, area, speed)
  - Hardened logic is less flexible

- **Going Wide**
  - Design complexity, area, and therefore power grows super-linearly
  - Need tools to ease user design complexity
  - CAD improvements for ultra-wide (512b – 5120b+) data paths
Programmability needed – what type?

- **CPU**
  - Fast programming turn-around
  - Many programmers
  - Very flexible
  - Power hungry
  - Limited speed for some applications

- **ASIC**
  - Power efficient; good speed
  - High NRE / long development time

- **FPGA**
  - “It was the best of times, it was the worst of times; it was the age of wisdom…”

© 2011 Altera Corporation—Public

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.
Answer: A little bit of this, a little bit of that

**CPU(s)**
- Very flexible
- Great for tasks that can take many cycles
- Can be hard, soft, or tightly coupled

**FPGA fabric**
- Wireline speed, bit manipulation, data-steering
- Flexibility, differentiation, in-field upgrades
- Volumes that don’t justify hardening

**Embedded Hardened Blocks**
- Standard interfaces and common functions
- Reduced power and area
40G Packet Processor

- Example system has 16 dual-core Task Processing Units
- 400MHz multi-threaded soft RISC CPU operation
  - EP4SGX230KF40C2 – medium density
- Programmable logic for 10GE (PCS/MACs), interconnect, and lookup engines
- Next gen will have hardened 10GE PCS
Examples where fabric vs CPU required

- Protocol origination/termination
  - Framing, Data scrambling, striping, CRC, FEC, ACK,…

- Data steering
  - Data switch, dynamic steering of packets to appropriate processor, control plane, or memory

- Data aggregation (joining of streams)
  - E.g. combining ten 40G streams per QoS requirements into single 400G stream
What to Harden?

- Embedded HardCopy is great, but what’s next?
  - Most suited to protocols and other logic near transceiver
  - Doesn’t solve problem of efficient core logic
What to Harden?

- Embedded HardCopy is great, but what’s next?
  - Most suited to protocols and other logic near transceiver
  - Doesn’t solve problem of efficient core logic
- What hard-logic can be put in core?
  - DSP is an example of success – general purpose and serves wide market (although not wireline)
  - CRC seemed like good idea, but how to generalize?
    - Different CRC polynomials
    - Different data widths of processing
    - 32b input easy enough that not worth it
  - Cross-bar / barrel shifter?
  - Building block that serves many areas?
  - Research area?
Wide data paths: example current cores

- **100Gbps Ethernet:**
  - 10 serdes @ 10.3125Gbps
  - 320b data bus @ 315 Mhz

- **150G Interlaken (typically used w/100GE):**
  - 24 serdes @ 6.25Gbps
  - 512b data bus @ ~300 Mhz

- **PCle gen3 x8**
  - 8 serdes @ 8Gbps
  - 256b data bus @ 250 Mhz
Going Wide – Where is the SOP?

- Most protocols have variable length packets
  - Start with SOP followed by data payload, ending with CRC
  - Granularity of packet length varies by protocol, 8b, 32b, 64b

- Prefer logic on natural word-size
  - Challenge when multiple words per parallel cycle
  - Difficult when SOP happens anywhere
  - Consider simple case of CRC
  - Harder still if multiple SOP per clock tick

64b x 1562 Mhz

384b x 260 Mhz
Concluding thoughts

- Current success result of past innovation
  - DSP and arithmetic blocks for hardened logic
  - Embedded HardCopy for protocol
  - Various low-power techniques such as back biasing

- Terabit will require future innovation

- Research Challenges
  - What general purpose hardened building blocks should be included and where?
  - How do we make the FPGA portion easier to develop?
  - How do we make FPGA architectures and CAD (User development, synthesis, P&R) more friendly to wide data-busses (512b – 5000+)?
Backup
SOP challenge harder as get wider

- 320b bus for Ethernet implies five SOP locations
  - Variable length packets imply SOP location constantly moves
  - Consider challenge of looking for Dest-MAC, offset from SOP

- 1024b bus implies 16 locations and two simultaneous SOPs on a single clock-tick
SOP anywhere solution (sort-of)

- SOP anywhere leads to added complexity
  - Looking for given field now depends on SOP position, e.g. MAC SRC
  - Significant additional barrel shifting required

- One alternative to SOP anywhere is to speed-things up and move SOP to the left:

<table>
<thead>
<tr>
<th>Data</th>
<th>Data</th>
<th>Data</th>
<th>Data</th>
<th>Data</th>
<th>Data</th>
<th>Data</th>
<th>SOP3</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>SOP2</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
</tr>
<tr>
<td>SOP1</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
<td>Data</td>
</tr>
</tbody>
</table>

  | SOP3 | Data |      |      |      |      |      |      |      |
  | Data | Data | Data | Data | Data | Data | X    | X    | X    |
  | SOP2 | Data | Data | Data | Data | Data | Data | Data | Data |
  | Data | X    | X    | X    | X    | X    | X    | X    | X    |
  | SOP1 | Data | Data | Data | Data | Data | Data | Data | Data |

- Challenge is that you have to run faster…
## 100GbE w/SOP at MSB position

<table>
<thead>
<tr>
<th>Width</th>
<th>Frequency</th>
<th>Over-clocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>64b</td>
<td>1562.5 Mhz</td>
<td>1562.5 Mhz</td>
</tr>
<tr>
<td>128b</td>
<td>781 Mhz</td>
<td>781 Mhz</td>
</tr>
<tr>
<td>256b</td>
<td>391 Mhz</td>
<td>441 Mhz (13%)</td>
</tr>
<tr>
<td>320b</td>
<td>312 Mhz</td>
<td>379 Mhz (18%)</td>
</tr>
<tr>
<td>384b</td>
<td>260 Mhz</td>
<td>321 Mhz (23%)</td>
</tr>
<tr>
<td>512b</td>
<td>195 Mhz</td>
<td>294 Mhz (51%)</td>
</tr>
<tr>
<td>768b</td>
<td>130 Mhz</td>
<td>213 Mhz (64%)</td>
</tr>
<tr>
<td>1024b</td>
<td>98 Mhz</td>
<td>168 Mhz (72%)</td>
</tr>
<tr>
<td>2048b</td>
<td>49 Mhz</td>
<td>149 Mhz (205%)</td>
</tr>
<tr>
<td>4096b</td>
<td>24 Mhz</td>
<td>149 Mhz (510%)</td>
</tr>
</tbody>
</table>

Table assumes 20B IPG + preamble, 0 inserted header bytes, 100GbE
Same issue applies for PCIe gen2 x8 and PCIe gen3 x8
# 400GbE w/SOP at MSB position

<table>
<thead>
<tr>
<th>Width</th>
<th>Frequency</th>
<th>Over-clocked</th>
</tr>
</thead>
<tbody>
<tr>
<td>64b</td>
<td>6,250 Mhz</td>
<td>6,250 Mhz</td>
</tr>
<tr>
<td>128b</td>
<td>3,128 Mhz</td>
<td>3,124 Mhz</td>
</tr>
<tr>
<td>256b</td>
<td>1,564 Mhz</td>
<td>1,764 Mhz (13%)</td>
</tr>
<tr>
<td>320b</td>
<td>1,248 Mhz</td>
<td>1,516 Mhz (18%)</td>
</tr>
<tr>
<td>384b</td>
<td>1,040 Mhz</td>
<td>1,284 Mhz (23%)</td>
</tr>
<tr>
<td>512b</td>
<td>780 Mhz</td>
<td>1,176 Mhz (51%)</td>
</tr>
<tr>
<td>768b</td>
<td>520 Mhz</td>
<td>852 Mhz (64%)</td>
</tr>
<tr>
<td>1024b</td>
<td>392 Mhz</td>
<td>672 Mhz (72%)</td>
</tr>
<tr>
<td>2048b</td>
<td>196 Mhz</td>
<td>596 Mhz (205%)</td>
</tr>
<tr>
<td>4096b</td>
<td>96 Mhz</td>
<td>596 Mhz (510%)</td>
</tr>
</tbody>
</table>

Table assumes 20B IPG + preamble, 0 inserted header bytes, 100GbE. Same issue applies for PCIe gen2 x8 and PCIe gen3 x8.