NoC with Near-Ideal Express Virtual Channels Using Global-Line Communication

Tushar Krishna¹, Amit Kumar¹, Patrick Chiang², Mattan Erez³, Li-Shiuan Peh¹

¹ Princeton University
² Oregon State University
³ University of Texas, Austin
The CMP era...

Global wires do not scale

Local communication with multi-hop networks
Not all interconnects are equal

- Conventional repeated RC wires
  - R. Ho [2001]
  - Latency several clock cycles across a chip (~3ns / 10mm)
  - High BW for short lengths

- On-Die Transmission Lines
  - K. Shepard [05,06], Ito [08]
  - Speed of light propagation (~100ps / 10mm)
  - Power and bandwidth density is poor

- Current Sensing/ Capacitive feed-forward
  - R. Ho [07], E. Mensink [07]
  - 5-10x improvement vs. conventional RC Wire (~500ps / 10mm)
  - BW density 2-4X lower (vs. short wires)

Single cycle global communication is possible

Trade-off latency vs. bandwidth
What if single cycle global communication is possible?

- Network-on-chip with hybrid interconnect

- Data plane
  - Multi-hop network
  - High bandwidth
  - Full-swing

- Control plane
  - Global lines (G-lines)
  - Ultra-low latency
  - Multi-drop

- Express virtual channels (EVCs) that rely on NOCHI
  - Critical flow control information is shared among routers using G-lines
  - Reduced buffering and power overhead in routers
S-CSMA Circuit Design

- 7 cores traversing 7mm
- Each TX using $C_{\text{CAP}}=300\text{fF}$
- Each Core RX with sense amplifier converting to digital
- Last RX uses amplitude detection to determine # of TXs transmitting concurrently
  - Uses Flash ADC (6 sense amplifiers with different ref. voltage)
Simulation Results

- Circuit simulation using 7 metal, 90nm 1.2V, CMOS process
- Pulse response along G-lines (up to 6 cores)
- Eye diagram at Flash ADC input
Express Virtual Channels*

- Virtual express lanes in the network
- Flits on these EVCs can bypass buffering and switch arbitration at intermediate routers
  - 5 stage Normal pipeline
  - 2 stage EVC pipeline

Express Virtual Channels (EVCs)*

- Virtual express lanes in the network
- Flits on these EVCs can bypass buffering and switch arbitration at intermediate routers
  - Normal Scenario: 15 cycles
  - Express virtual channel (EVC): 9 cycles

G-line EVCs

- Use G-lines for global flow-control information
  - G-line control plane and conventional data plane

- Down-stream node broadcasts buffer/VC availability
  - Single cycle to all up-stream nodes
  - All nodes have up-to-date information

- Up-stream nodes request resources in 1 cycle
  - Each down-stream node has 1 buffer-request and 1 VC-request G-line
  - Shared line enables single-cycle count of requests
    - No need to identify requesting nodes, just the count
  - Uses S-CSMA circuit
Setup

- 7x7 packet-switched mesh network
- Two G-lines per node & direction for buffer/VC request
The signaling mechanism

Number of granted requests are sent to the upstream on the data plane
Benefits of G-line EVCs

• Instantaneous global information
  – Aggressive buffer management
  – Original EVC reserved buffers for signal traversal time

• Broadcast medium
  – Enables flexible, dynamic EVCs of any length
  – Original EVC limited by signaling cost
    • Partition VCs into k-hop bins
    • Limits EVCs to short lengths (< 3 hops)
Network evaluation (1)

- For the same number of buffers per port

![Graph showing latency vs. injected traffic for EVC and NOCHI]

- 9.4% reduction at low load
- 44% reduction near saturation
- Can handle more load before hitting the saturation point

Tornado Traffic
For the same saturation throughput

<table>
<thead>
<tr>
<th>Network power for original EVC</th>
<th>Network power for G-line EVC</th>
<th>Net power reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>47.99W</td>
<td>43.76W</td>
<td>4.23 W (8.8%)</td>
</tr>
</tbody>
</table>

* Power numbers based on extrapolation of Intel Polaris 80-core network
Conclusions

• Effective long-range communication is possible
  – 1 cycle cross-chip with reasonable power
  – Lower overall bandwidth
• NOCHI utilizes low-latency control plane and high-bandwidth data plane
  – Single cycle, multi-drop, broadcast for control
  – Full-swing multi-hop network for data
• The advantages of EVCs (latency and power) are enhanced by using NOCHI
  – EVCs of arbitrary lengths allowed
  – No conservative buffer management
  – Dynamic binding of VCs to different EVC lengths
Thanks
Backup
EVCs in Action

Flit

On/Off

Lookahead (to set up switch)

On/Off

Lookahead (to set up switch)

On/Off

Buffer

VC

Crossbar