Variable-Width Datapath for On-Chip Network Static Power Reduction

George Michelogiannakis, John Shalf
Postdoctoral Research Fellow
Computer Architecture Laboratory
Lawrence Berkeley National Laboratory
In a Nutshell

- Leakage power is an increasing problem in future or near threshold voltage (NTV) technologies

- Leakage power can be important even at high network loads

- This work proposes variable-width datapaths
  - Parts of channels, buffers, and crossbars can be activated on demand

- We demonstrate an average of 33% total power reduction with PARSEC benchmarks
Today’s Menu

- Leakage power / motivation
- Related work
- Variable-width datapaths
- Results
- Conclusions
Leakage Power Contribution


Subthreshold Leakage at NTV

NTV operation reduces total power, improves energy efficiency
Subthreshold leakage power is substantial portion of the total

Assumes 20% leakage power at 100% Vdd
Increasing Variations

Near-threshold voltage (NTV) design — Opportunities and challenges. DAC 2012
Applications Load Network Unevenly

“Fine-grained bandwidth adaptivity in networks-on-chip using bidirectional channels”. NOCS 2012
Today’s Menu

- Leakage power / motivation
- Related work
- Variable-width datapaths
- Results
- Conclusions
High threshold voltage “sleep switch” transistor

Savings when sleep time enough to overcome energy overheads

“MP3: Minimizing Performance Penalty for Power-gating of Clos Network-on-Chip”. HPCA 2014
Drowsy SRAMs

- Put SRAM lines into low-power (low voltage) “drowsy” mode
  - Preserves data
- Faster activation than power-gated SRAMs (1-2 cycles)
  - Higher leakage current while drowsy. Higher activation penalty

CatNap: Multiple Networks

High traffic area

Multinets cannot share resources (e.g., channels) in low traffic regions

Optimal decisions at injection are a challenge

“Catnap: Energy Proportional Multiple Network-on-Chip”. ISCA 2013
Today’s Menu

- Leakage power / motivation
- Related work
- Variable-width datapaths
- Results
- Conclusions
Variable-Width Channels

Same bisection bandwidth
If flits from any channel lane can choose any VC, that necessitates multiplexers.

We map channel lanes to use only a subset of VCs (1:1 with equal VCs and lanes).

“Adding slow-silent virtual channels for low-power on-chip networks”. NOCS 2008
Crossbar Gating

“Segment gating for static energy reduction in networks-on-chip”. NoCArc 2009
Activation Mechanism

- Flits winning switch allocation (SA) activate in the next router:
  - Output channels and switch lanes (3 cycles)
  - Input buffers (1 cycle with drowsy SRAMs)
    - No false activations
  - With the below 4-stage router pipeline, no activation stalls
ABN switch allocators: \((\text{Inputs} \times \text{VCs}) \times (\text{Outputs} \times \text{ChanLanes})\)

- As long as ChanLanes no greater than VCs, switch allocator no more complex than VC allocator
- If VC and switch allocators in different pipeline stages, router cycle time does not extend

VC allocators consume 2-10mW and occupy 5000um

- Both very small percentages of the router
- Therefore increase of switch allocator’s cost insignificant

“Allocator implementations for network-on-chip routers”. SC 2009
Today’s Menu

- Leakage power / motivation
- Related work
- Variable-width datapaths
- Results
- Conclusions
Methodology

- Booksim network simulator
- 8x8 Mesh. DOR

We compare:
- **Single-lane**: Single-lane power-gated network
- **ABN**: Flits choose a lane based on their output VC
- **Multinets**: Multiple power-gated subnetworks

- Router pipeline previously presented
- 2 VCs as baseline. Normalize for buffer size by adjusting VCs

Activation and deactivation delays (65nm at 1GHz):
- Channel and crossbar activation delay: 3 cycles
- Channel and crossbar activation wait: 15 cycles
- Channel and crossbar deactivation wait: 6 cycles
- Buffer (VC) deactivation wait: 3 cycles
Two Subnetworks/Lanes. Static Power

8x8 mesh. DOR. UR traffic

Static power (W)

Injection rate (request packets/cycle * 1000)

Single lane
ABN
MultiNets

UR worst case for ABNs
ABNs better powers down resources at high loads
Two Subnetworks/Lanes. Dynamic Power

8x8 mesh. DOR. UR traffic

- Single lane
- ABN
- MultiNets

UR worst case for ABNs
Multinets takes advantage of lower-radix switches
Multinets cannot make perfect injection decisions or use resources in another subnetwork after injection to combat transient imbalance.
Two ABN lanes and two multinet subnetworks

Percentage total power reduction (%)

- Blackscholes
- Canneal
- Dedup
- Ferret
- Fluidanimate
- VIPS
- X264
- Average

Low  Medium  High
Four ABN lanes and four multinet subnetworks
Effects of transient imbalance in multinets are intensified
Today’s Menu

- Leakage power / motivation
- Related work
- Variable-width datapaths
- Results
- Conclusions
Conclusions

- Leakage power is a growing concern in future technologies

- Dividing datapaths in lanes provides more flexibility than multi-network approaches
  - But there are tradeoffs

- Using drowsy SRAMs allows hiding the activation delay without false activations
  - Can change with shallow router pipelines

- We demonstrate an average of 33% total power reduction with PARSEC benchmarks
Questions?

Acknowledgment: D.O.E. office of science