Communication
The Next Resource War

Simon Moore & Daniel Greenfield

SLIP – Invited Talk, April 6th 2008

UNIVERSITY OF CAMBRIDGE
Computer Laboratory
Computer Architecture Group

Overview

Background
Rent's Rule for NoCs
Communication in Algorithms
Conclusions & Research Questions

Computation vs. Communication

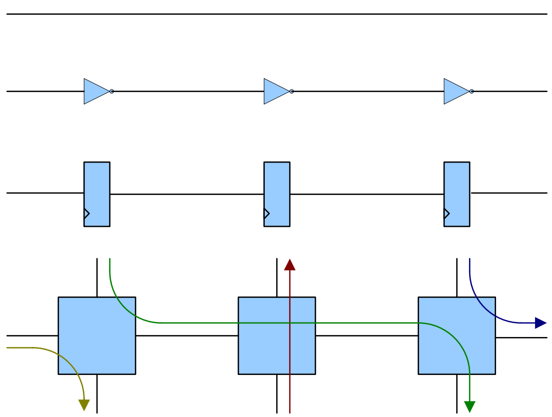
- Relative power consumed

technology node	130nm CMOS	50nm CMOS
transfer 32b across chip	20 ALU ops	57 ALU ops
transfer 32b off-chip	260 ALU ops	1300 ALU ops

When did global wire scaling stop?

- Simple global interconnect has hardly improved in 30 years!
  - chip area has changed little since the birth of the microprocessor
  - thinner wires don't help and newer materials are a one-off trick
- It's only now that it has started to hurt

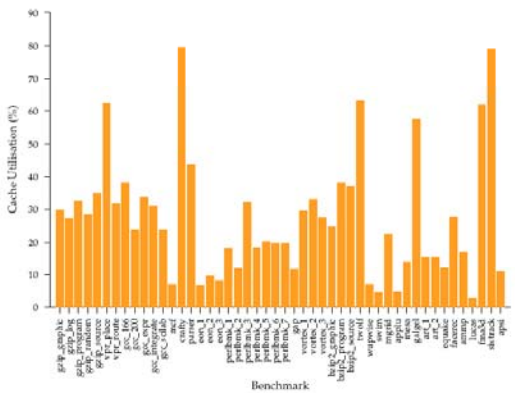
Virtualization of Interconnect



Locality of Data

- The main weapon to minimise communication
- Current approaches:
  - caching
    - relies on statistical properties of temporal and address locality to provide hardware support
  - scratch pad memories
    - places the burden on the programmer

Level-2 unified cache utilisation



From James Srinivasan, University of Cambridge, Computer Laboratory

The problem with caches

- Often 80% of the cache holds dead data
- That's a huge waste of transistors
- **We need to be smarter about exploiting locality**

Overview

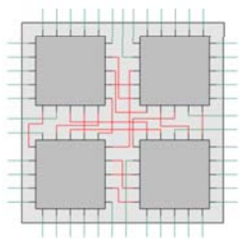
Background

Rent's Rule for NoCs

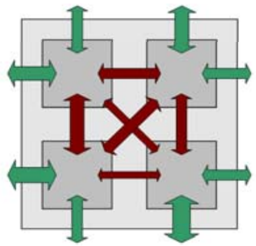
Communication in Algorithms

Conclusions & Research Questions

A New Rent's Rule



$T = kG^p$



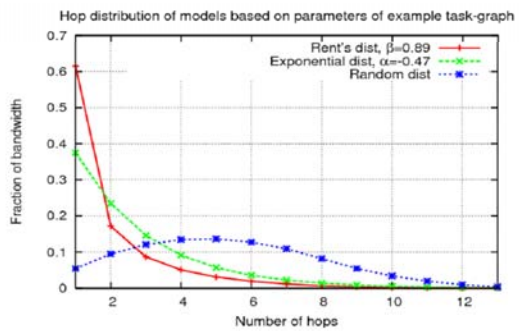
$B = bN^\beta$

Why Expect This?

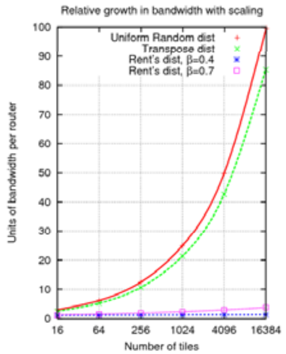
Domain to minimize	Wires	NoC
Delay	Wire delay	NoC latency (& congestion)
Congestion	Wire-density	Cross-sectional BW
Power	Wire buffering & length	Hop-length & router-utilisation

- BUT Needs
  - Topology supporting multi-scale locality
  - Mapping with locality as implicit or explicit goal
  - **Communication graphs with multi-scale / fractal locality properties**

Why Care: Statistics



Why Care: Scaling



- Common synthetic traffic models do not scale
  – Independent of topology

Overview

Background

Rent's Rule for NoCs

Communication in Algorithms

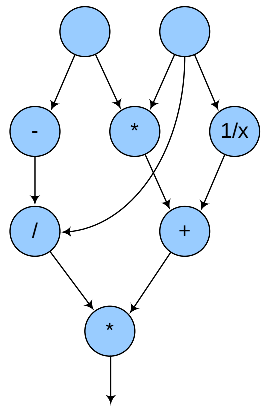
Conclusions & Research Questions

Communication Constraints in SW

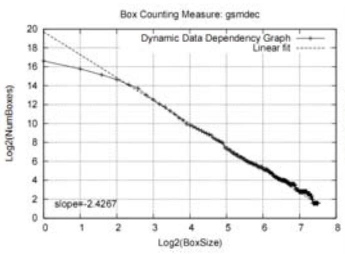
- Chip Multiprocessors (CMP) on NoC
  – Different to multi-chip multiprocessors
  – Much greater on-chip bandwidth
  – Lower latencies
  – Supports fine-grain parallelism
- Communication in algorithms
  – Poor understanding of communication locality
  – How much locality can be extracted / exploited?
  – What fundamental properties do they possess?
  – Can we model the locality?

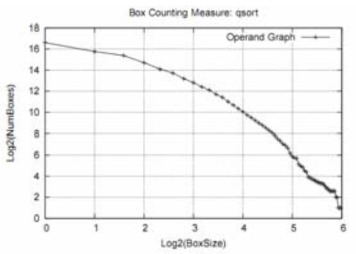
Software Graphs

- Dynamic data dependency graph
  – graph representation of computation data dependencies
- Assumes perfect oracle of control-flow decisions
- Edges
  – communication via RF/caches/external-mem/virtual-mem/etc
- Graph distance vs. instruction distance

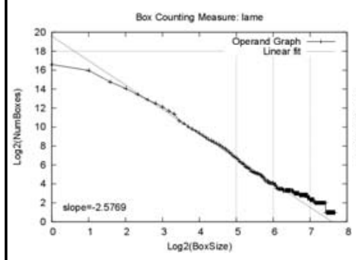


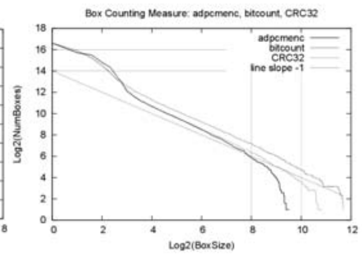
Fractal Communication in SW



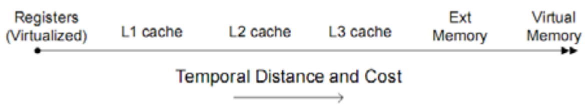


Fractal Communication in SW



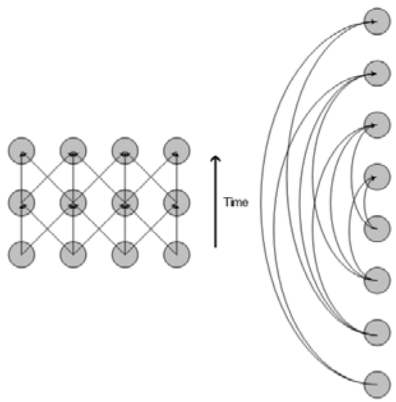


Memory as Temporal Interconnect

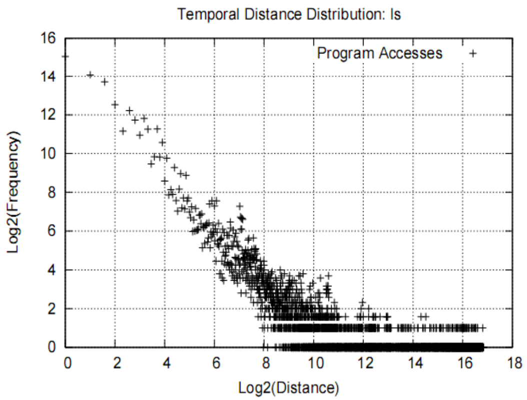


- Memory as wires
  - Register files connecting instruction output to input
- Memory as LUT
  - Replace functions with mux'ed data values
- Memory as switch, part of network

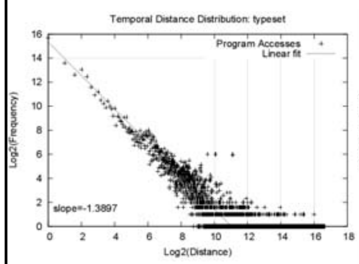
Spatio-Temporal View

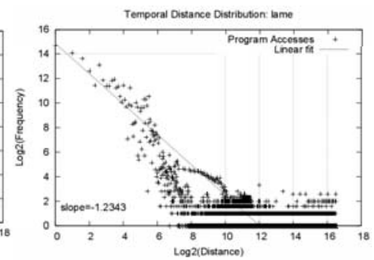


Temporal Interconnect: Rentian?



Temporal Interconnect: Rentian?





Overview

Background

Rent's Rule for NoCs

Communication in Algorithms

Conclusions & Research Questions

Conclusions and Research Questions

- Networks-on-chip transforms physical interconnect into virtual interconnect
- Adding virtualisation/indirection resolves many problems in computer science, but how do we maximise the benefits?
  - \+ Higher utilisation
  - \+ Specialised interconnect
  - \+ Higher abstraction / modular composition
  - \- Latency
  - \- Scheduling
  - \- Area

Conclusions and Research Questions

- Software exhibits fractal locality
  - Supports requirements for Rentian statistics
  - Can we exploit this behaviour?
  - Can we automatically reduce communication complexity/dimensionality?
  - How tight are the dimensionality constraints on communication statistics?

Conclusions and Research Questions

- Memory as temporal interconnect
  - Similarities to spatial interconnect / switch
  - Distance distributions appear Rentian?
  - Can we leverage our statistical models to design better temporal interconnect?
- Unification of views
  - Data is routed in space and time
  - What new techniques can we develop by unifying spatial and temporal communication?

Contact Details

Computer Architecture group web page:
http://www.cl.cam.ac.uk/research/comparch

Email:
simon.moore@cl.cam.ac.uk
daniel.greenfield@cl.cam.ac.uk