UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Department of Computer Engineering

Analyzing topology parameters for achieving energy-efficient k-ary n-cubes

HiPINEB 2018

Francisco José Andújar Muñoz et al.

Outline

Outline

Introduction

- When the network is designed, we must search a trade-off between:
  - Network performance.
  - Economic cost $\Rightarrow$ Deployment cost $+$ Exploitation cost.
    - The exploitation cost greatly depends on energy consumption.
    - Greater performance $\Rightarrow$ Greater energy consumption $\Rightarrow$ Greater cost
- We must seek the most energy-efficient network:
  - The network that requires lower energy for doing the same job.
  - However, not every performance penalty is acceptable.

Energy-efficiency in torus topology

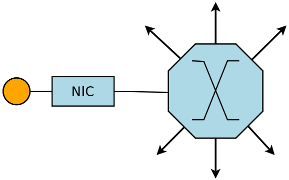
- Torus topology is widely used in supercomputers.
- What torus configuration is more energy-efficient? Given:
  - A fixed number of nodes.
  - The same bisection bandwidth in each compared network.

Energy-efficiency in torus topology

- Torus topology is widely used in supercomputers.
- What torus configuration is more energy-efficient? Given:
  - A fixed number of nodes.
  - The same bisection bandwidth in each compared network.
- Two main possible configurations:
  - High-dimensional networks with a high number of low-degree switches.
  - Low-dimensional network with a low number of high-degree switches using link trunking.

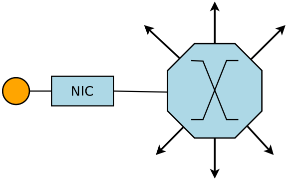
64-node network example

- Configuration A:
  - 4x4x4 3D torus
  - No trunk links
  - 7-port switches
  - 64 routers.
  - Average distance: 3
  - Network diameter: 6
  - 448 network ports

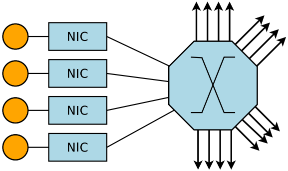


64-node network example

- Configuration A:
  - 4x4x4 3D torus
  - No trunk links
  - 7-port switches
  - 64 routers.
  - Average distance: 3
  - Network diameter: 6
  - 448 network ports



- Configuration B:
  - 4x4 2D torus
  - 4 ports per trunk link
  - 20-port switches
  - 16 routers
  - Average distance: 2
  - Network diameter: 4
  - 320 network ports



64-node network example

- Configuration B has:
  - the same bisection bandwidth as A.
  - 71.4% ports of configuration A...
  - ... and then, lower power consumption.

64-node network example

- Configuration B has:
  - the same bisection bandwidth as A.
  - 71.4% ports of configuration A...
  - ... and then, lower power consumption.
- But $E = P * t$:

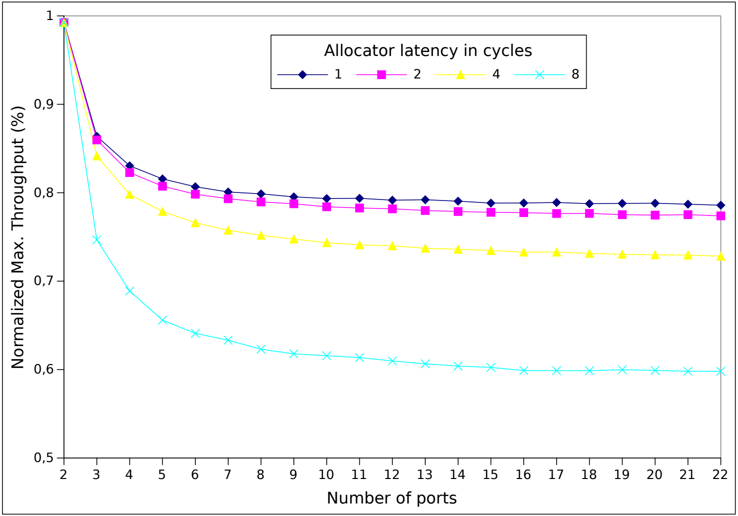
64-node network example

- Configuration B has:
  - the same bisection bandwidth as A.
  - 71.4% ports of configuration A...
  - ... and then, lower power consumption.
- But $E = P * t$:
  - B has lower diameter and average distance than A.
  - B has high-degree switches.

64-node network example

- Configuration B has:
  - the same bisection bandwidth as A.
  - 71.4% ports of configuration A...
  - ... and then, lower power consumption.
- But $E = P * t$:
  - B has lower diameter and average distance than A.
  - B has high-degree switches.
  - The number of ports affects the switch allocator performance:
    - Requires bigger (and slower) round-robin arbiter.
    - The allocator performance slightly decreases.

Allocator performance



Energy vs performance

- What network...
  - ...achieves the highest performance?
  - ...is the most energy-efficient?
- If the most energy-efficient network has not the greatest performance, is the performance penalty acceptable?

Energy vs performance

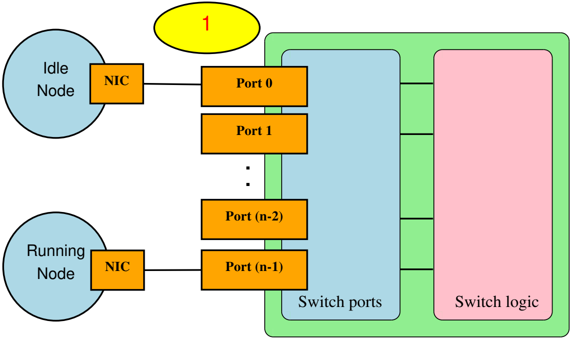
- What network...
  - ...achieves the highest performance?
  - ...is the most energy-efficient?
- If the most energy-efficient network has not the greatest performance, is the performance penalty acceptable?
- To answer these questions, we need to:
  - Define a power consumption model.
  - Evaluate the networks by simulation.
  - Use the simulation results in the power model.

Outline

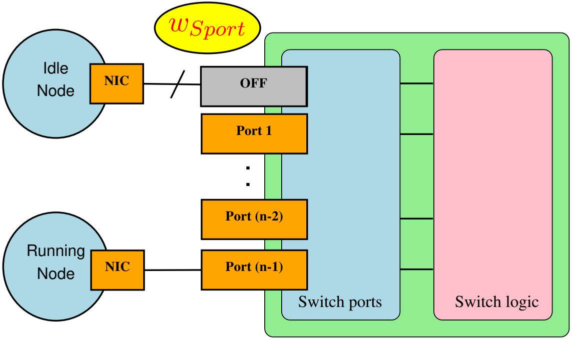
Initial hypotheses

- The switch power consumption increases linearly with the number of ports.
- Two states for the switch ports: *wake-up* (or *turned on*) and *sleep* (or *turned off*).
- Two states for the compute nodes: *running* or *idle*.
- Color guide:
  - Topology parameters.
  - Power model parameters.
  - Simulation statistics.
  - Metric estimated by the power model.

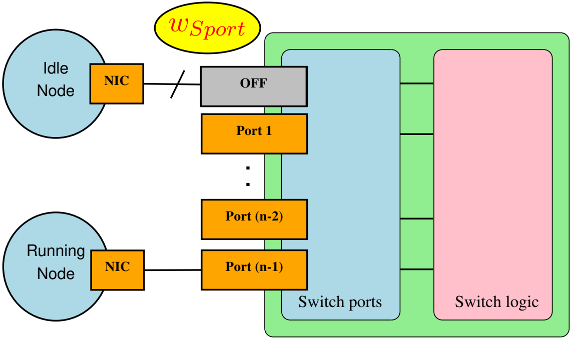
Port power model



Port power model

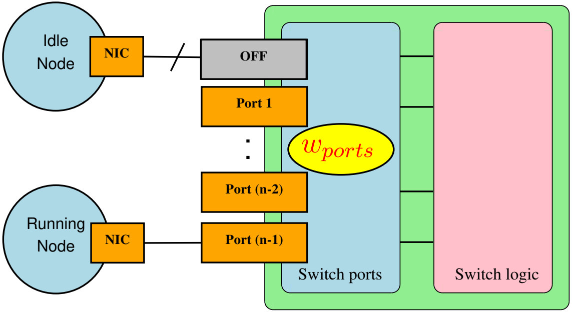


Port power model

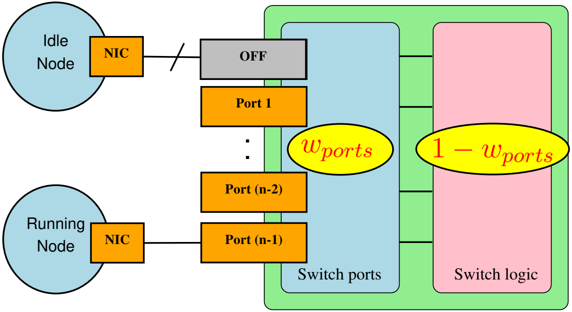


$$W^{s}_{ports} = w_{Sports} + (1 - w_{Sports}) \times U_{ports}$$

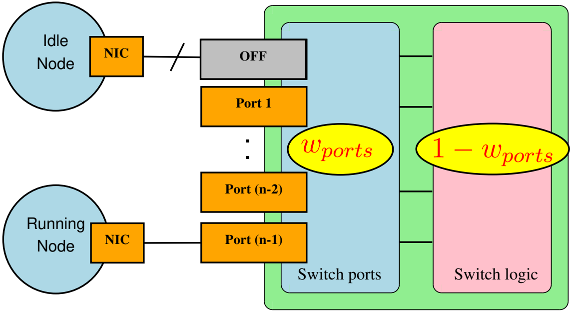
Switch power model



Switch power model

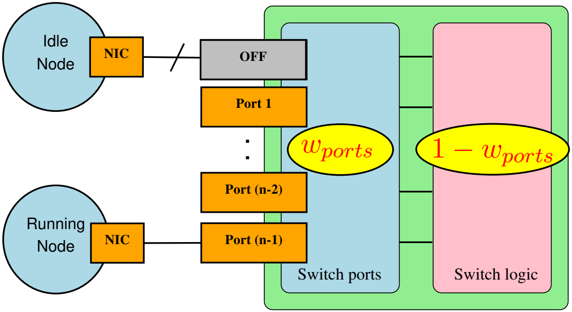


Switch power model



$$W_{sw}^{s} = (1 - w_{ports}) + w_{ports} \times W_{ports}^{s}$$

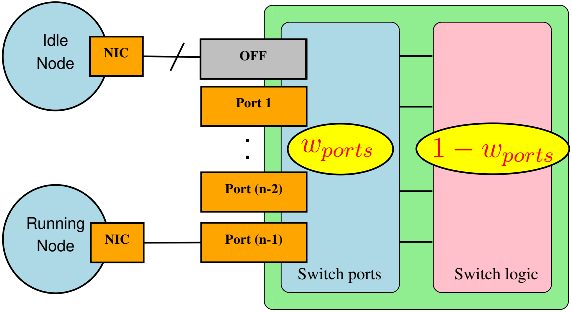
Switch power model



$$W^s_{sw} = (1 - w_{ports}) + w_{ports} \times W^s_{ports}$$

$$W_{net} = \frac{1}{N_{sw}} \sum_{i=1}^{N_{sw}} W^i_{sw}$$

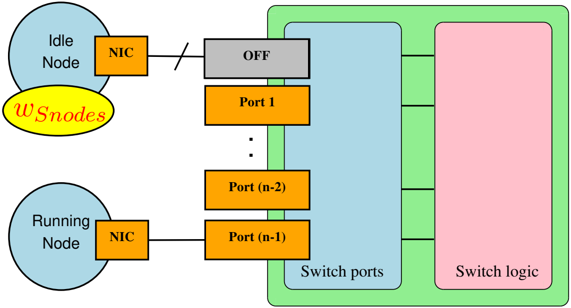
Switch power model



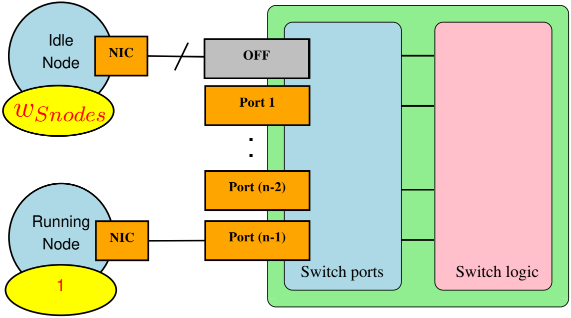
$$W_{sw}^{s} = (1 - w_{ports}) + w_{ports} \times W_{ports}^{s}$$

$$W_{net} = \frac{1}{N_{sw}} \sum_{i=1}^{N_{sw}} W_{sw}^{i} \times \frac{N_{ports}}{REF_{ports}} \times \frac{N_{sw}}{REF_{sw}}$$

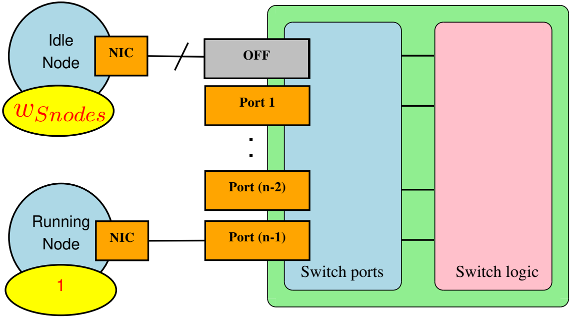
Compute node power model



Compute node power model

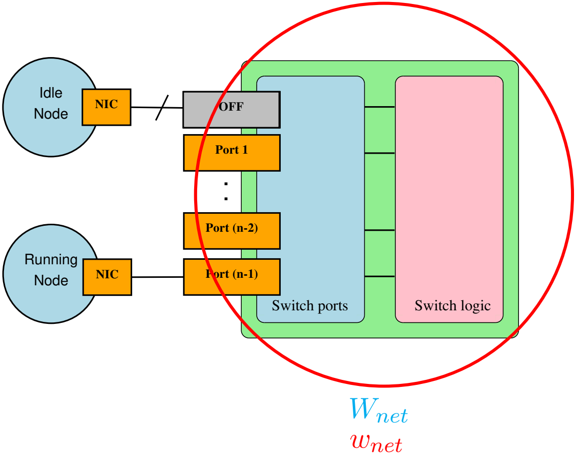


Compute node power model

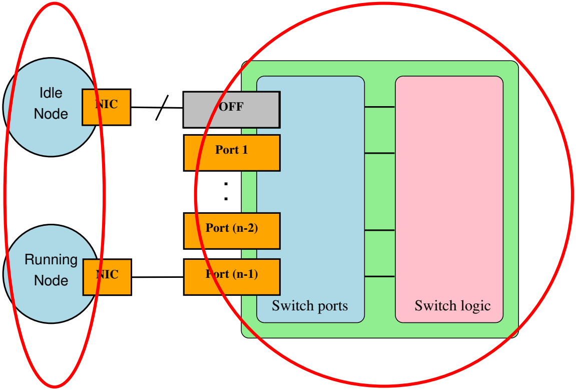


$$W_{nodes} = w_{Snodes} + (1 - w_{Snodes}) \times U_{cpu}$$

Cluster power model



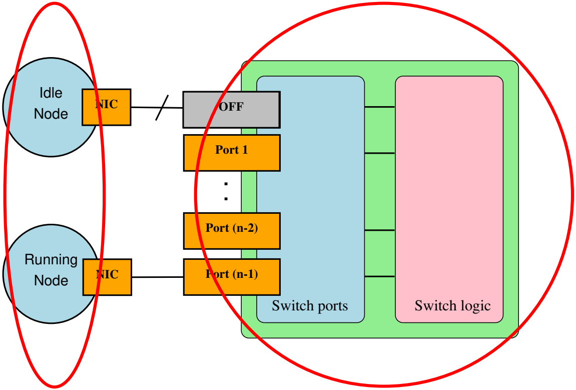
Cluster power model



W_{nodes}
$(1 - w_{net})$

W_{net}
w_{net}

Cluster power model

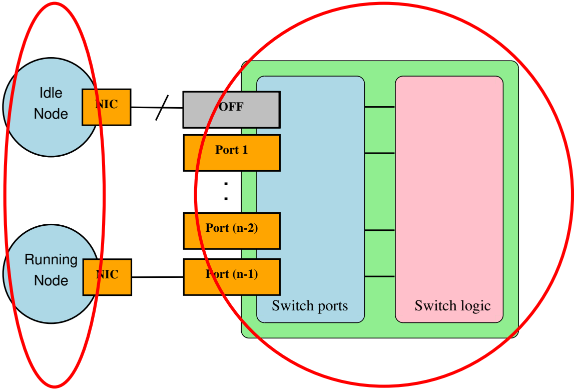


W_{nodes}
$(1 - w_{net})$

W_{net}
w_{net}

$$W_{cluster} = w_{net} \times W_{net} + (1 - w_{net}) \times W_{nodes}$$

Cluster power model



W_{nodes}
$(1 - w_{net})$

W_{net}
w_{net}

$$W_{cluster} = w_{net} \times W_{net} + (1 - w_{net}) \times W_{nodes}$$

$$E_{cluster} = W_{cluster} \times RunTime$$

Power model parametrization

Parameter	Value
w_{Sport}	0.1
w_{ports}	0.65
w_{net}	0.15
w_{Snodes}	Variable

Outline

Switch architecture

- IQ switches
- Virtual cut-through
- Credit flow-control
- 3-stage allocator (based on Blue Gene allocator)
  - Round-Robin arbiter latency logarithmically increases with the number of ports
- Routing algorithm: fully-adaptive routing (Duato's protocol).
- Port bandwidth: 10 GBytes/s
- A trunk link:
  - Comprises several independent ports
  - Each port transmits independent packets
  - Power saving: the number of wake-up ports depends on trunk link utilization

Workload model

- VEF traces:
  - Traces obtained from MPI applications
  - Self-related traces:
    - Each communication depends on a previous communication
    - The changes in the network are reflected in the execution time
- Selected applications:
  - NAMD (smtv benchmark)
  - HPPC MPI Random Access
  - Graph500 benchmark
- Each trace has 512 MPI tasks

Case Studies

	64 nodes		256 nodes	
Topology	3D	2D	4D	3D
Dimensions	4x4x4	4x4	4x4x4x4	4x4x4
Num. Ports	7	20	9	28
Allocator latency	3	5	4	5
Port Aggregation	1	4	1	4
Num. Switches	64	16	256	64
Network Ports	448	320	2304	1792
Port Ratio	–	0.714	–	0.777

Outline

Introduction

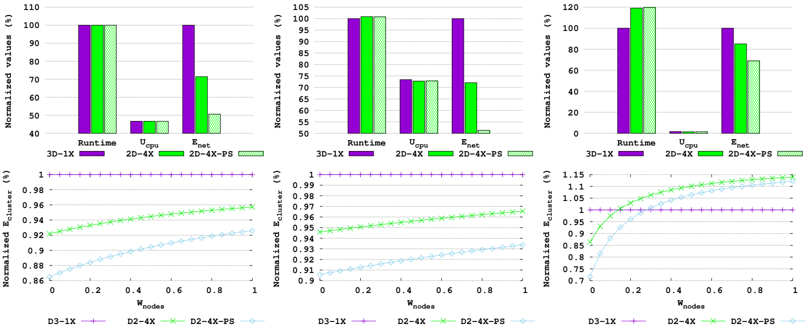
Power model

System model

Evaluation

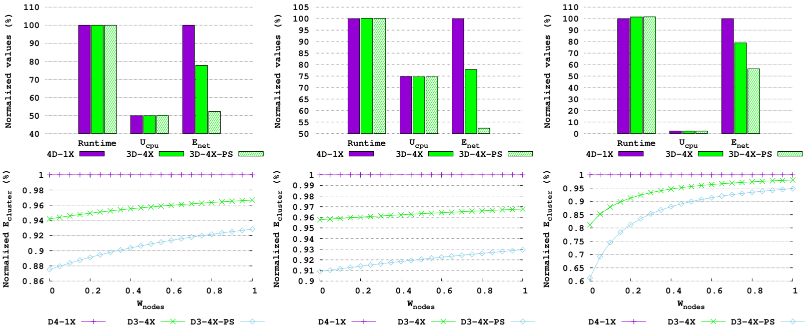
Conclusions and future work

64-node networks



(a) NAMD (b) MPI Random Access (c) Graph500

256-node networks



(d) NAMD (e) MPI Random Access (f) Graph500

Outline

Conclusions

- Under low and medium traffic loads:
  - No differences in performance.
  - Trunk-link torus is more energy-efficient.
- Under high traffic loads:
  - Trunk-link torus has a significant performance penalty.
  - High-dimensional torus are more energy-efficient...
  - ... unless the compute nodes are very energy-proportional.
- In the trunk-link torus, the power-saving mechanism has:
  - No significant performance penalty.
  - Lower energy consumption.

Future Work

- Evaluation using more MPI applications.
- Evaluate more topologies:
  - Fat-tree
  - Dragonfly
- Simple trace scheduler to evaluate the topologies under a large set of applications instead of a single application.

Department of Computer Engineering

Analyzing topology parameters for achieving energy-efficient k-ary n-cubes

HiPINEB 2018

Francisco José Andújar Muñoz et al.
fraanmu1@upv.es