Full fat or skinny?

How do you like your compute nodes?

Mike Giles

Oxford e-Research Centre

Introduction

Two different approaches to HPC:

- small number of fat nodes

  future systems at Oak Ridge and Lawrence Livermore will have multiple IBM Power9 CPUs and NVIDIA Volta GPUs, connected by multiple Infiniband network adapters

  future system at Argonne will have nodes based on Intel Xeon Phi with silicon photonic networking

- large number of thin nodes

  similar to IBM Blue Gene, could use large number of SoC (System-on-Chip) micro-servers with low-cost 10 GigE networking

Which is best?

How can we model/assess the pros and cons?

Outline

- construct a simple model of a scientific computation

  (computation, local data transfer, remote data transfer)

- construct a simple model of its energy cost

  (flops, memory and network BW, node and memory power)

- look at energy / power data for real hardware to draw some tentative conclusions

Computational model

- G grid points, split amongst N processes (1 per node)
- d-dimensional layout, so $(G/N)^{1/d}$ points in each direction
  $\Longrightarrow O((G/N)^{-1/d})$ boundary points per interior point
- f flops per point

  m bytes per point

  b bytes per point to/from node memory

  $a\,(G/N)^{-1/d}$ bytes per point to/from other processes
- representative values for one timestep of a CFD code:

  $d = 3$

  $m = 4000$

  $f = 800$

  $b = 10000$

  $a = 50000$

Energy cost

- P power consumption per node

  P_M power consumption per byte of memory

  F flop/s

  B byte/s memory BW

  A byte/s network BW

- execution time

$$\frac{G}{N}\left(\frac{f}{F}+\frac{b}{B}+\frac{a}{A}\left(\frac{G}{N}\right)^{-1/d}\right)$$

- execution energy

$$\frac{G}{N}\left(\frac{f}{F}+\frac{b}{B}+\frac{a}{A}\left(\frac{G}{N}\right)^{-1/d}\right)(N\,P+G\,m\,P_M)$$

Energy cost

Dividing by G and re-arranging gives the following energy per grid point

$$E = \left(\frac{f}{\widehat{F}} + \frac{b}{\widehat{B}} + \frac{a}{\widehat{A}}\left(\frac{G}{N}\right)^{-1/d}\right)\left(1 + m\,\widehat{P}_M\left(\frac{G}{N}\right)\right)$$

where

$$\begin{aligned} \widehat{F} &= F/P \text{ (flops/J)} \\ \widehat{B} &= B/P \text{ (bytes/J)} \\ \widehat{A} &= A/P \text{ (bytes/J)} \\ \widehat{P}_M &= P_M/P \end{aligned}$$

This gives $E \approx \dfrac{f}{\widehat{F}} + \dfrac{b}{\widehat{B}}$ provided

$$\left(\frac{a}{\widehat{A}} \Big/ \left(\frac{f}{\widehat{F}} + \frac{b}{\widehat{B}}\right)\right)^d \ll \frac{G}{N} \ll \frac{1}{m\,\widehat{P}_M}$$

Energy cost

Minimising the extra cost due to the networking time and the memory,

$$\frac{a}{\widehat{A}}\left(\frac{G}{N}\right)^{-1/d} + \left(\frac{f}{\widehat{F}} + \frac{b}{\widehat{B}}\right) m \widehat{P}_M \left(\frac{G}{N}\right)$$

gives a near-optimal value

$$\left(\frac{G}{N}\right)_{opt} = \left\{\frac{1}{d\, m \widehat{P}_M}\left(\frac{a}{\widehat{A}}\right)\left(\frac{f}{\widehat{F}} + \frac{b}{\widehat{B}}\right)^{-1}\right\}^{d/(d+1)}$$

Note: only near-optimal because this neglects the cross-product term $\left(\frac{a}{\widehat{A}}\right)\left(\frac{G}{N}\right)^{-1/d} \times m \widehat{P}_M \left(\frac{G}{N}\right)$.

Hardware

We consider 3 different node configurations:

- Node 1:
  - 2 NVIDIA K80 GPUs
  - 2 Intel 10-core Xeon E5-2650 CPUs
  - 2 Mellanox ConnectX-4 dual-port adapters (each 2×100Gb/s)
  - GDDR5 + DDR4 memory
- Node 2:
  - 2 Intel 10-core Xeon E5-2650 CPUs
  - 2 Mellanox ConnectX-3 dual-port adapters (each 2×40Gb/s)
  - DDR4 memory
- Node 3:
  - 1 Intel 8-core Xeon D-1540 SoC CPU (built-in 2×10Gb/s Ethernet)
  - DDR4 memory

Hardware

quantity (units)	node 1	node 2	node 3
P (W)	$2\times300 + 2\times100 + 2\times15$	$2\times100 + 2\times10$	45
P_M (W/GB)	1.0	0.4	0.4
F (GFlop/s)	4000	500	80
B (GB/s)	1000	100	30
A (GB/s)	50	20	2.5
$\widehat{P}_M$ (1/byte)	1.2×10^{-12}	1.8×10^{-12}	9×10^{-12}
$\widehat{F}$ (flop/J)	4.8×10^{9}	2.3×10^{9}	1.8×10^{9}
$\widehat{B}$ (byte/J)	1.2×10^{9}	0.45×10^{9}	0.67×10^{9}
$\widehat{A}$ (byte/J)	60×10^{6}	91×10^{6}	56×10^{6}

For the CFD application, $f \approx b$, so it's bandwidth-limited, and best choice is node with lowest $\widehat{B}$, provided not constrained by memory cost or network bandwidth

Hardware

For the CFD application, the lower and upper bounds for G/N given by network bandwidth and memory power, respectively, are:

	node 1	node 2	node 3
lower, $\left((a/\widehat{A})\,/\left(f/\widehat{F}+b/\widehat{B}\right)\right)^d$	9×10^5	1.5×10^4	2×10^5
upper, $1/(m\,\widehat{P}_M)$	2.1×10^8	1.4×10^8	2.8×10^7

The optimal values for G/N and $m\,G/N$ (the memory per node) are:

	node 1	node 2	node 3
G/N	24×10^6	6.2×10^7	3.6×10^6
$m\,G/N$	95 GB	25 GB	14 GB

For node 1, 48 GB of GDDR5 memory is sub-optimal, and networking time increases energy consumption by 40% (assuming no overlap with compute)

Alternative energy model

The previous model assumes fixed power consumption by processor, memory, networking, regardless of what they are doing.

This is a gross idealisation – manufacturers have put considerable effort into reducing power consumption when not doing anything

(Also, increasingly programmers are getting option to adjust clocks in different parts of the system as appropriate to maximise power efficiency)

Energy cost

- $\underline{P}$ baseline power consumption per node
  $\overline{P}$ additional peak power consumption per node
  $\underline{P}_M$ baseline power consumption per byte of memory
  $\overline{P}_M$ additional peak power consumption per byte of memory

- execution time

$$\frac{G}{N}\left(\frac{f}{F}+\frac{b}{B}+\frac{a}{A}\left(\frac{G}{N}\right)^{-1/d}\right)$$

- execution energy

$$\frac{G}{N}\left\{\left(\frac{f}{F}+\frac{b}{B}+\frac{a}{A}\left(\frac{G}{N}\right)^{-1/d}\right)(N\,\underline{P}+G\,m\,\underline{P}_M)+\frac{f}{F}\,N\,\overline{P}+\frac{b}{B}\,G\,m\,\overline{P}_M\right\}$$

More complicated, but I don't think the conclusions would change much

Conclusions

- fairly simple modelling gives insight into issue of fat vs. thin nodes
- if not constrained by memory power or network bandwidth, best to use node with best flops/watt or bandwidth/watt
- fat nodes need very high networking bandwidth and/or much more memory

Notes:

- next-generation NVIDIA Pascal GPUs will have up to 64 GB of very efficient stacked memory – 3× bandwidth and 3× energy efficiency
- new Intel Xeon Phis will also have very efficient stacked memory, and future generation will have integrated photonic networking