How Efficient Can We Be?:
Bounds on Algorithm Energy Consumption

Andrew Gearhart

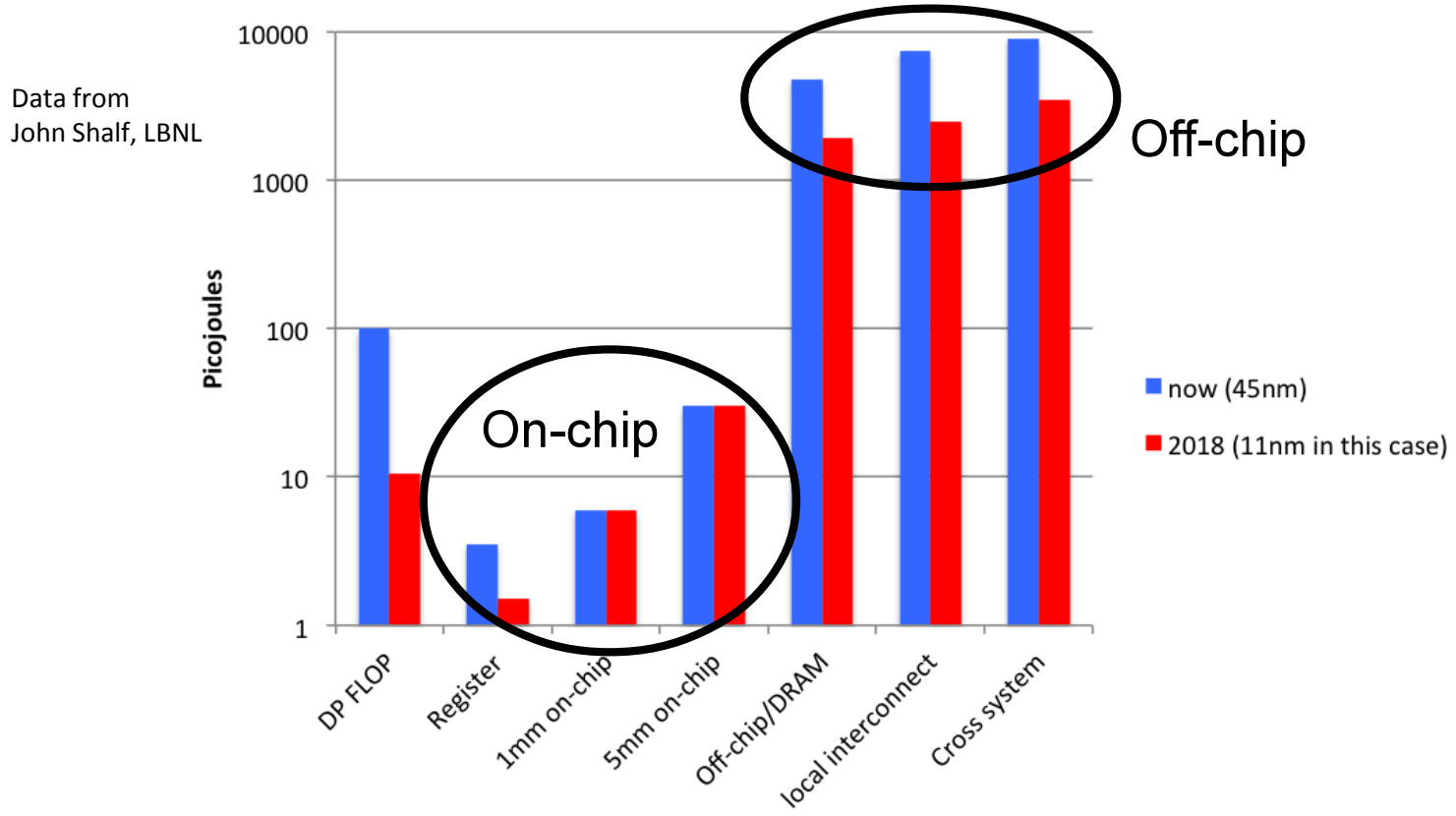
Relation to ASPIRE

- ASPIRE ("Algorithms and Specializers for Provably-optimal Implementations with Resiliency and Efficiency") -> recall Krste's talk
- "Provably-optimal" is the focus of this work
- Software and hardware design use feedback to "cotune" compute kernel energy efficiency

Previous Work: Communication Lower Bounds

- Bounds on bandwidth and number of messages for most of dense and sparse linear algebra
- Bounds for homo and heterogeneous machine models (i.e. GPU/CPU)
- Led to the development of many "communication-optimal" algorithms

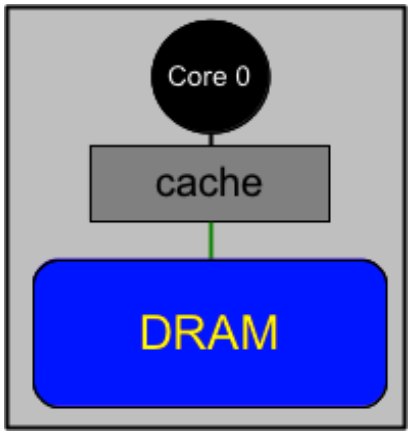
Communication is energy inefficient!



- On-chip/Off-chip gap isn't going to improve much

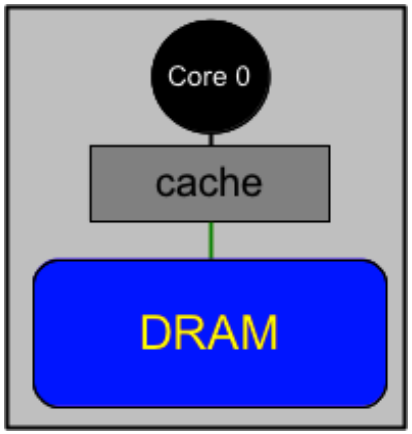
Communication is energy inefficient!

- **Communication lower bounds + machine models = lower bounds on energy**
- Machine models can be simple:

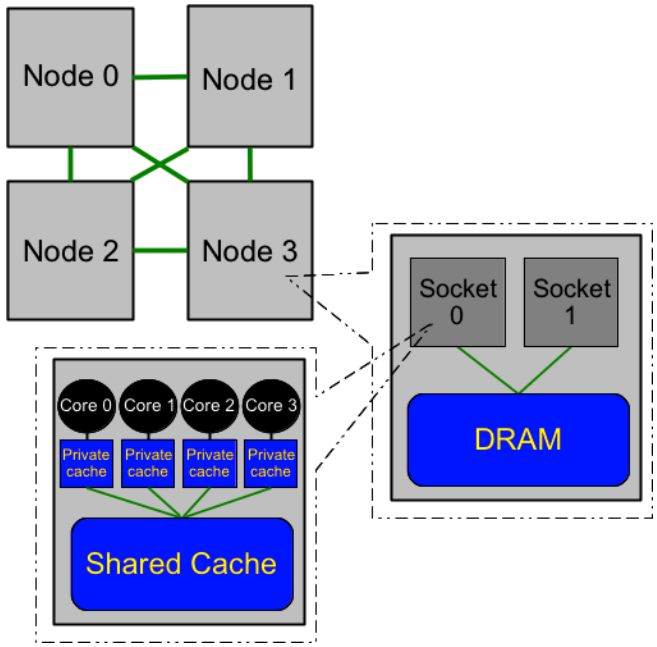


Communication is energy inefficient!

- **Communication lower bounds + machine models = lower bounds on energy**
- Machine models can be simple:



- Or more complicated…



Runtime and Energy Models

- Currently simple linear expressions that include bandwidth (W) and per transfer (S) costs -> link to communication bounds

Runtime and Energy Models

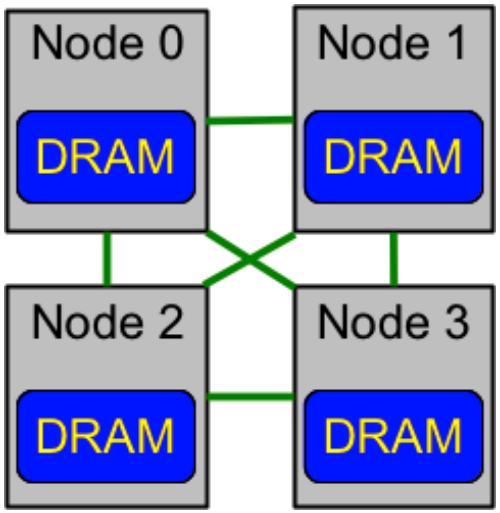
- Currently simple linear expressions that include bandwidth (**W**) and per transfer (**S**) costs -> link to communication bounds
- Models are based on a set of hardware and algorithmic parameters:

$$T = \gamma_t F + \beta_t W + \alpha_t S$$

$$E = p(\gamma_e F + \beta_e W + \alpha_e S + \delta_e MT + \epsilon_e T)$$

Example: 2.5D Matrix-Matrix Multiplication (GEMM)

- 2.5D algorithm calculates GEMM by replicating input data to reduce communication



$$T_{2.5DMM}(n,p,M) = \frac{\gamma_t n^3}{p} + \frac{\beta_t n^3}{M^{1/2}p} + \frac{\alpha_t n^3}{mM^{1/2}p} \quad (9)$$

$$\begin{aligned} E_{2.5DMM}(n,p,M) = & \quad (\gamma_e + \gamma_t\epsilon_e)n^3 \\ & + \left((\beta_e + \beta_t\epsilon_e) + \frac{(\alpha_e+\alpha_t\epsilon_e)}{m}\right)\frac{n^3}{M^{\frac{1}{2}}} \\ & + \quad \delta_e\gamma_t M n^3 \\ & + \quad \left(\delta_e\beta_t + \frac{\delta_e\alpha_t}{m}\right) M^{\frac{1}{2}} n^3. \quad (10) \end{aligned}$$

Energy Bounds

- Have energy bounds for GEMM, matrix-vector multiplication, Strassen's matrix multiplication, FFT, n-body, LU
- What are energy-optimal machine parameters for a given problem?
- **ASPIRE Open House on the 5th floor of Soda Hall!!!!**