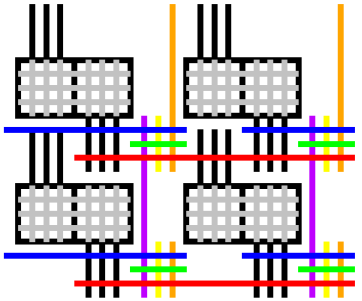
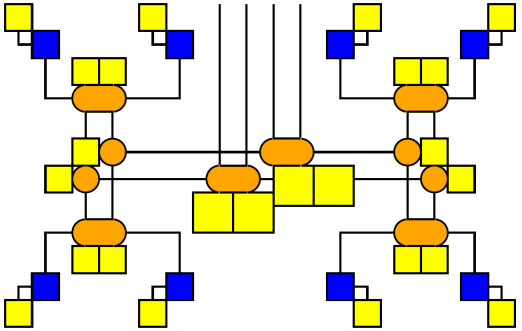




Location, Location, Location:
The Role of Spatial Locality in Asymptotic Energy Minimization

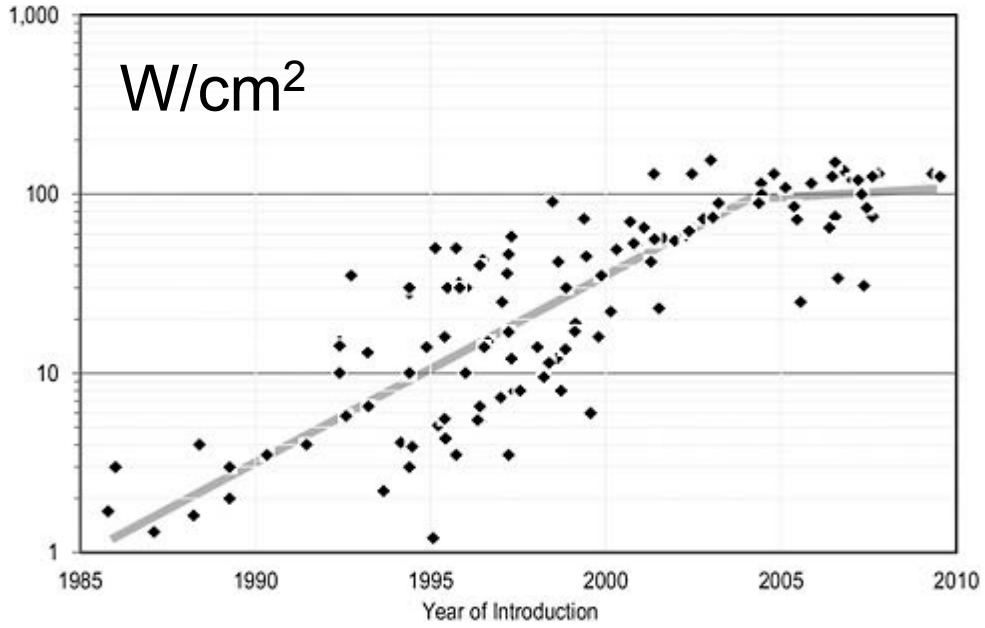
André DeHon

andre@seas.upenn.edu

Penn

Question

- Are FPGAs more or less **energy** efficient than processors?



The Future of Computing Performance: Game Over or Next Level?
National Academy Press, 2011

Importance:

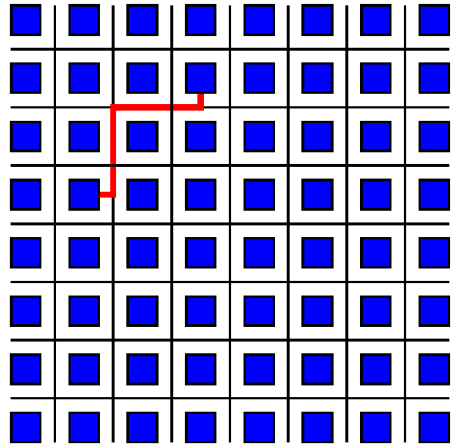
- Computation/chip limited by energy-density **not** transistor capacity → dark silicon
- Energy efficiency determines battery life

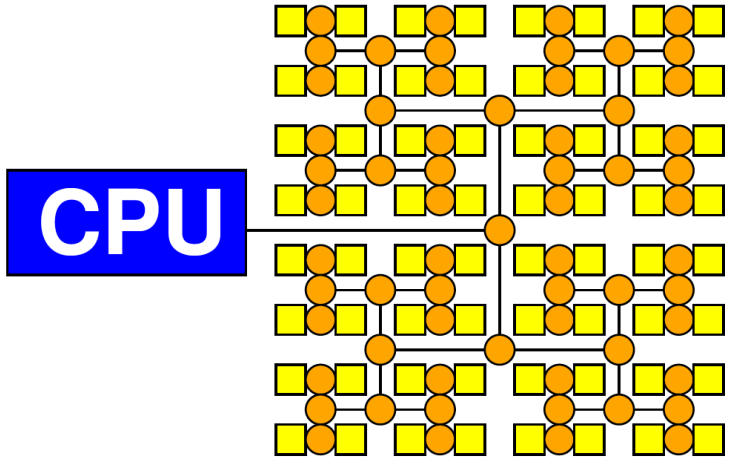
Findings

- Oversimplifed soundbyte:
  - FPGAs use less energy than processors
- More precisely:
  - There is an asymptotic advantage to
    - Parallel, spatial evaluation
    - Over sequential evaluation on central processor
  - When
    - There is any spatial locality (Rent $p<1.0$)
    - There is uniform average gate activity
- With "enough" spatial locality ($p<0.5$), FPGAs use $\Theta(N)$ energy

Intuition

- Given a good spatial layout
  - It is cheaper to transmit the result of a gate to its well-placed consumers
    - average wire length $O(N^{p-0.5}) \leq O(N^{0.5})$
      - $O(1)$ for $p<0.5$
  - Than to
    - Fetch inputs from a large central memory
      - $O(N^{0.5})$



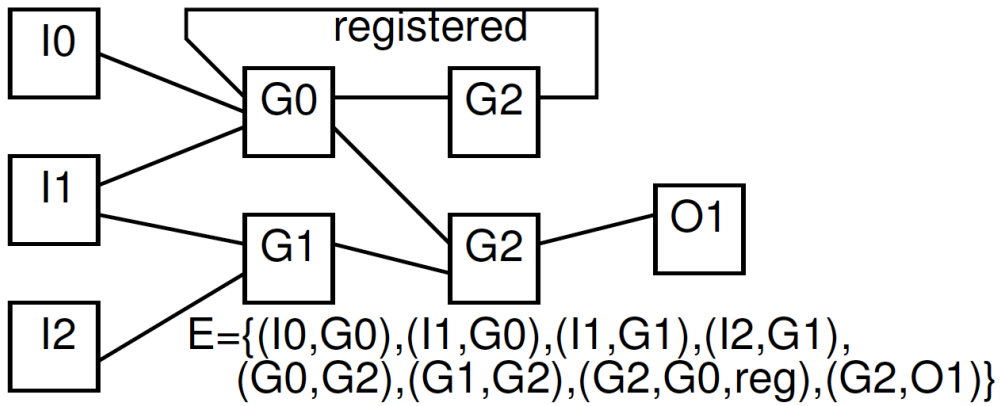


Outline

- Model
- Central Processor
- Spatial Locality
  - Description
  - Data
  - Instruction
- Asymptotic Results
- Unanswered

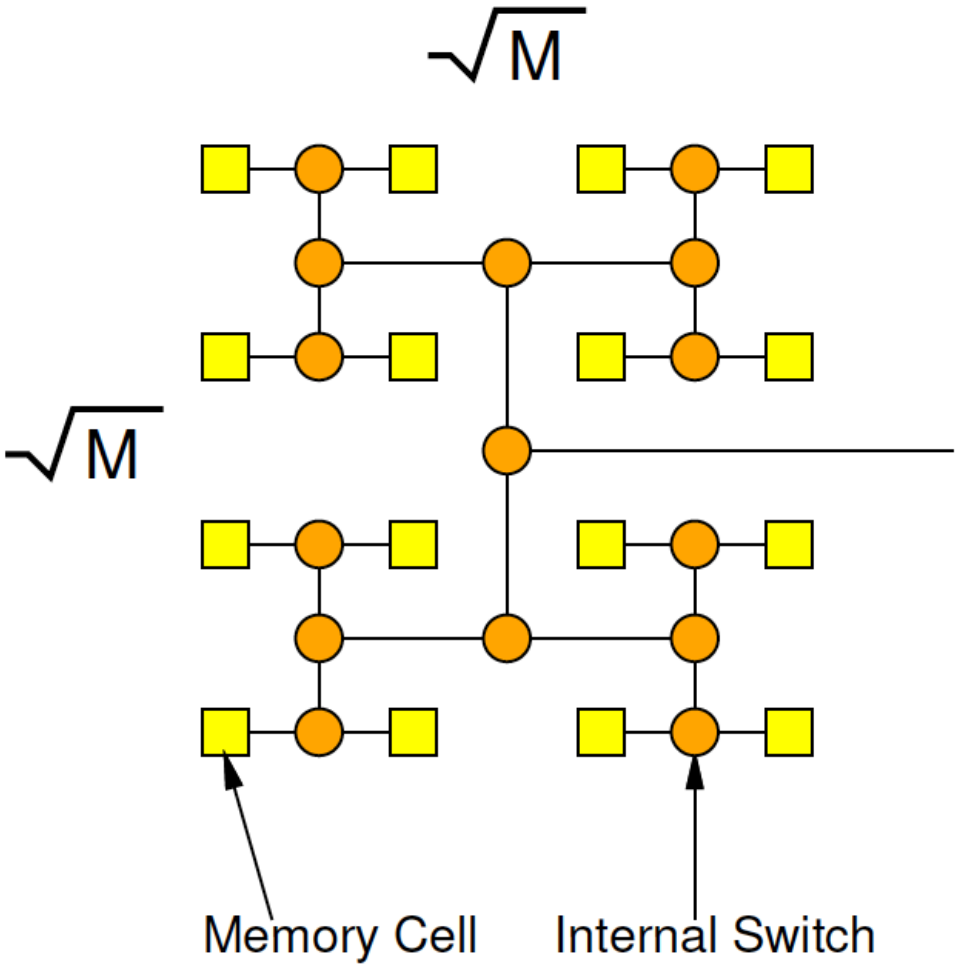
Gate Array Evaluation Model

- For each "circuit"
  - Evaluate N k-input gates (k constant, e.g. 4)
- Assume
  - Must evaluate every gate every time
  - Every gate active (may switch)



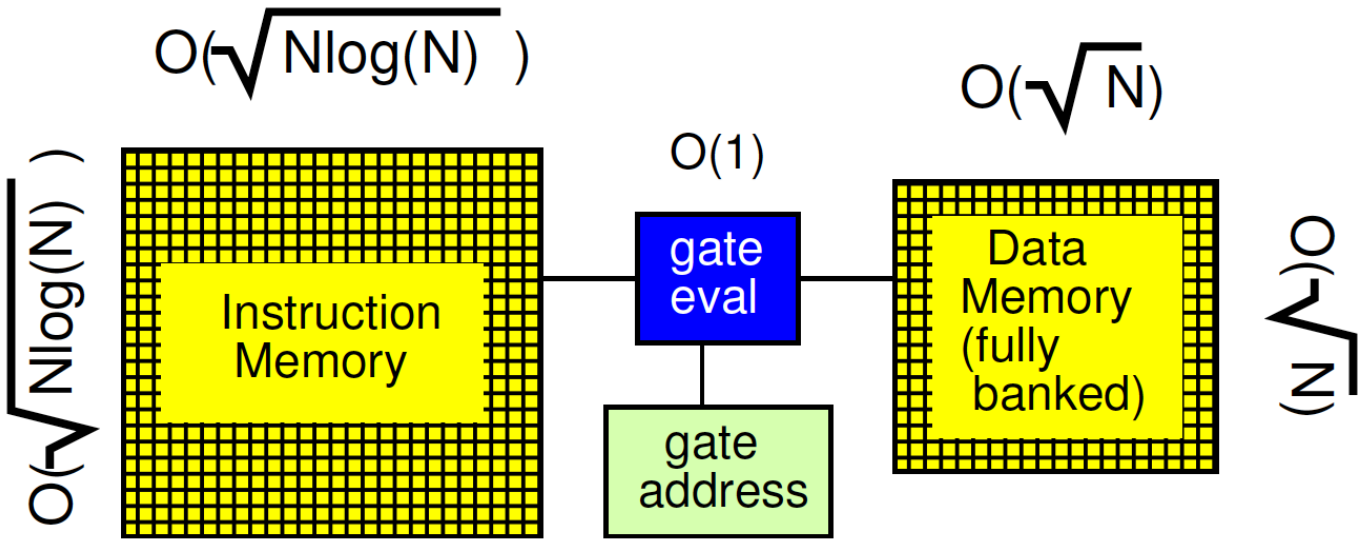
Fully Banked Memory

- Only activate path to leaf
- $O(M^{0.5})$ wire length
- Random access must send address
  - $O(M^{0.5}\log(M))$
- Sequential access avoid address per bit
  - $O(M^{0.5})$ per bit



Central Processor

- Each instruction specifies source nodes so is O(log(N)) long
  - $O(N^{1.5}\log^{1.5}(N))$
- Read/write O(N) bits
  - $O(N^{1.5}\log(N))$



Problem 1: Description Locality

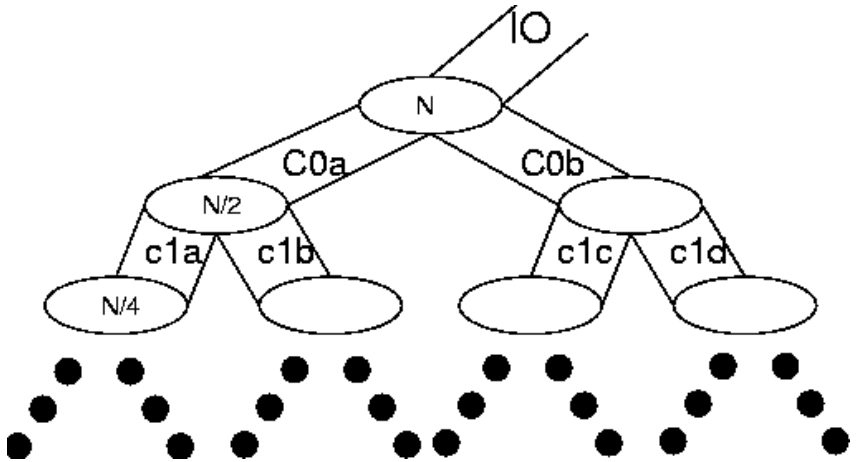
- Costs O(Nlog(N)) for description since can assume any input comes from anywhere

Spatial Locality

- If we place a circuit well,
  - Most of the inputs can be "close"
- More formally: Rent's Rule
  - If we recursively bisect a graph, attempting to minimize the cut size, we typically get:

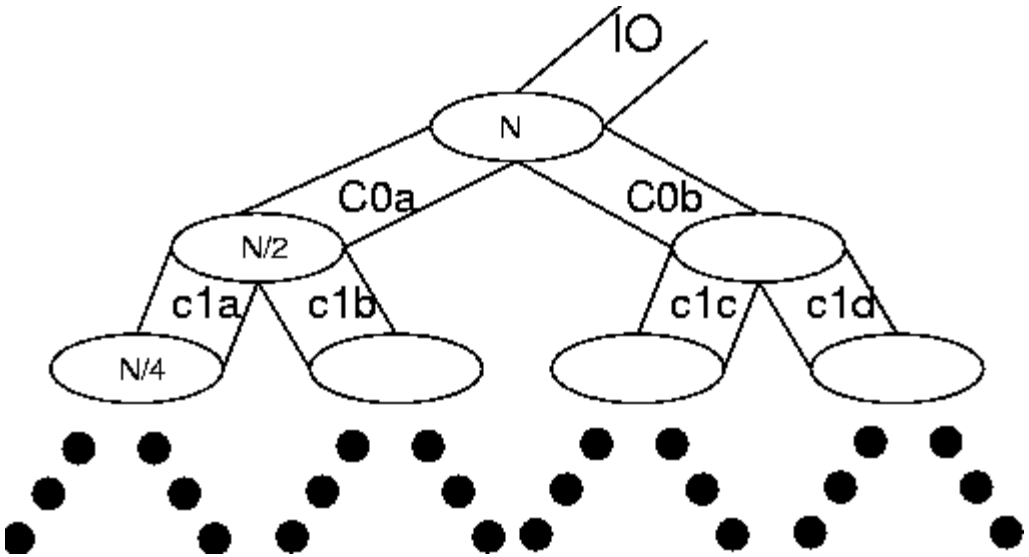
$$IO = c\ N^p$$

  - $0 \le p \le 1$
  - $p \le 1$ means many inputs come from within a partition



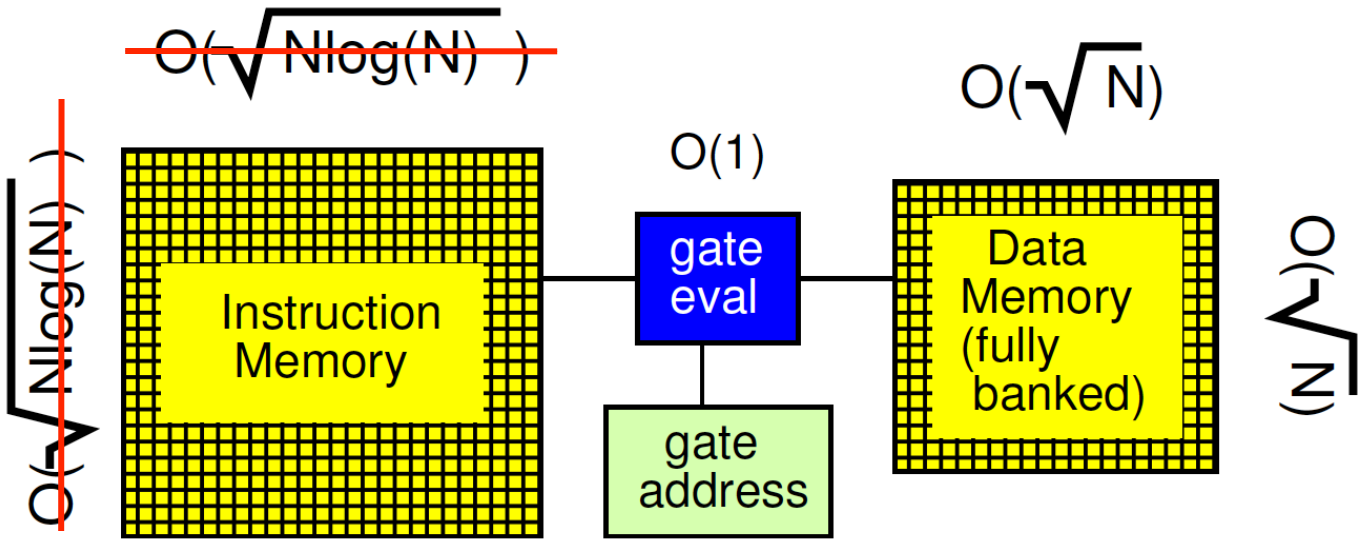
Description Lemma

- If p<1.0, can describe computation with O(N) memory.
  - If something close by, only need to use bits proportional to subtree height



Central Processor with Description Locality

- p<1.0, total instruction bits are O(N)
  - $O(N^{1.5})$
- Read/write O(N) bits
  - $O(N^{1.5}\log(N))$

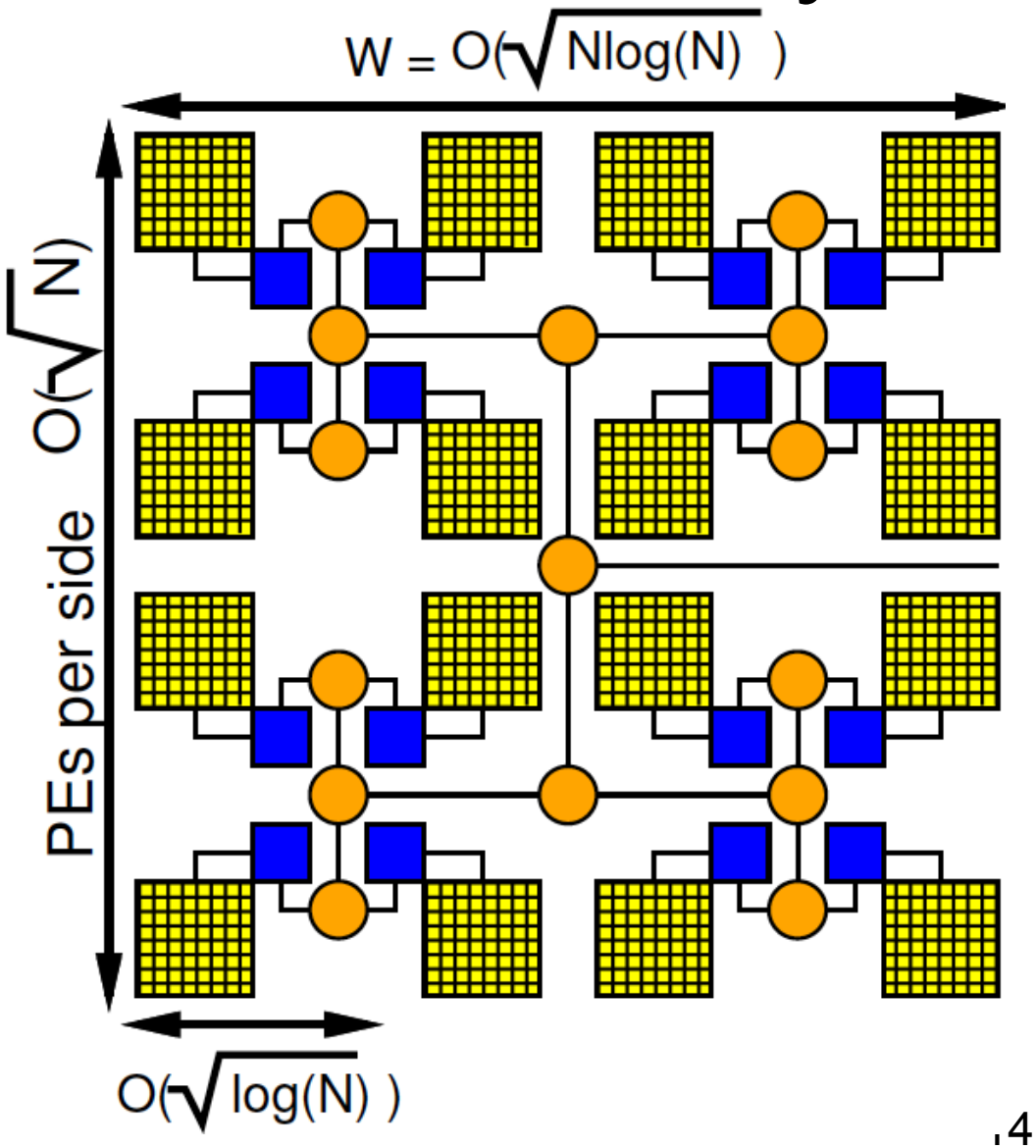


Problem 2: Data Locality

- Must pay $O(N^{0.5})$ for every read since data must be moved in and out of memory.

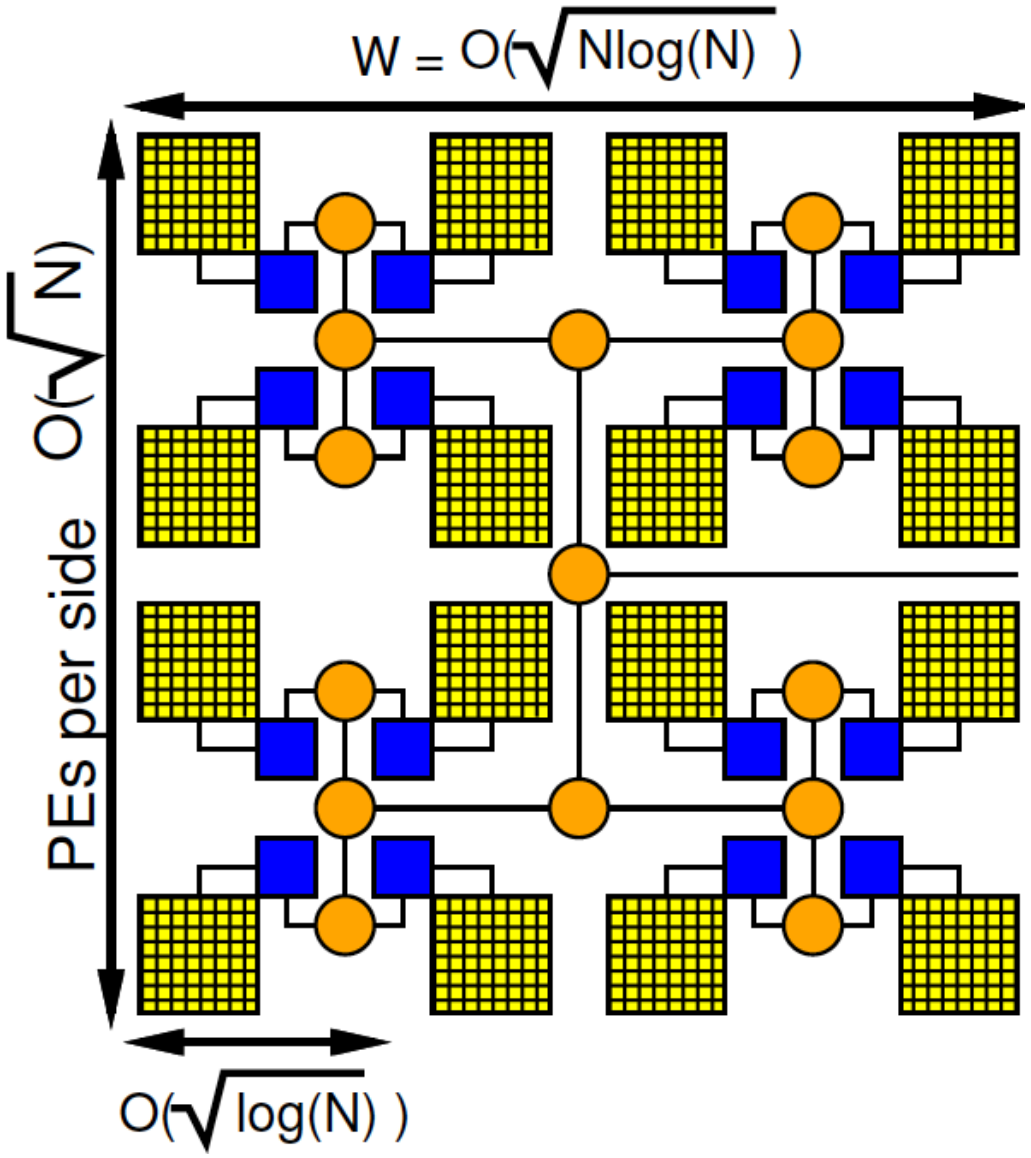
Sequential with Data Locality

- Store data at endpoints
- Send through network from producer to consumer
- Store location at leaves – O(log(N))
- Build H-Tree to keep area to O(Nlog(N))



Sequential with Data Locality

- Area = O(Nlog(N))
- Sending addresses log(N) bits at top
- Signals lower O(1)
- Only send a few over top $O(N^p)$
- $O((\log^{1.5}N)N^{p+0.5})$ for p>0.5
- **Cheaper to send where needed than to central location.**

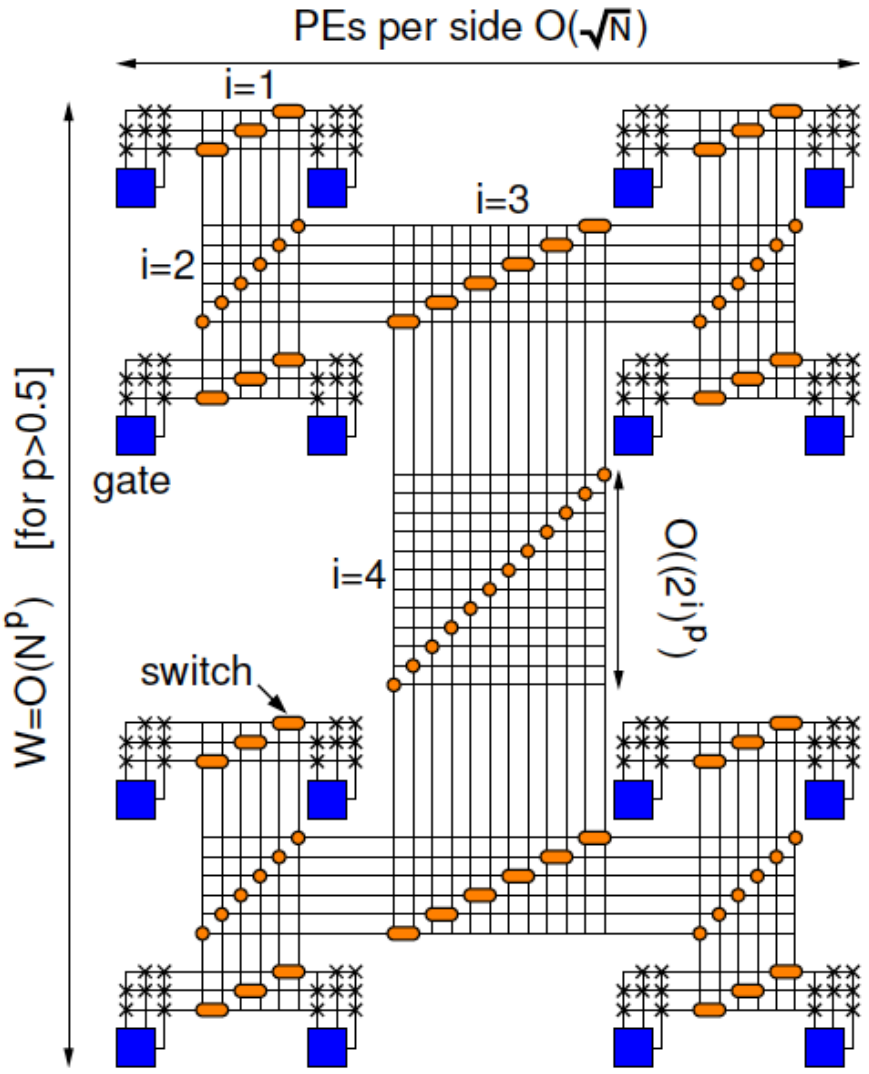


Problem 3:

- Multiply energy by O(log(N)) to send an address up the tree

Fully Spatial (FPGA)

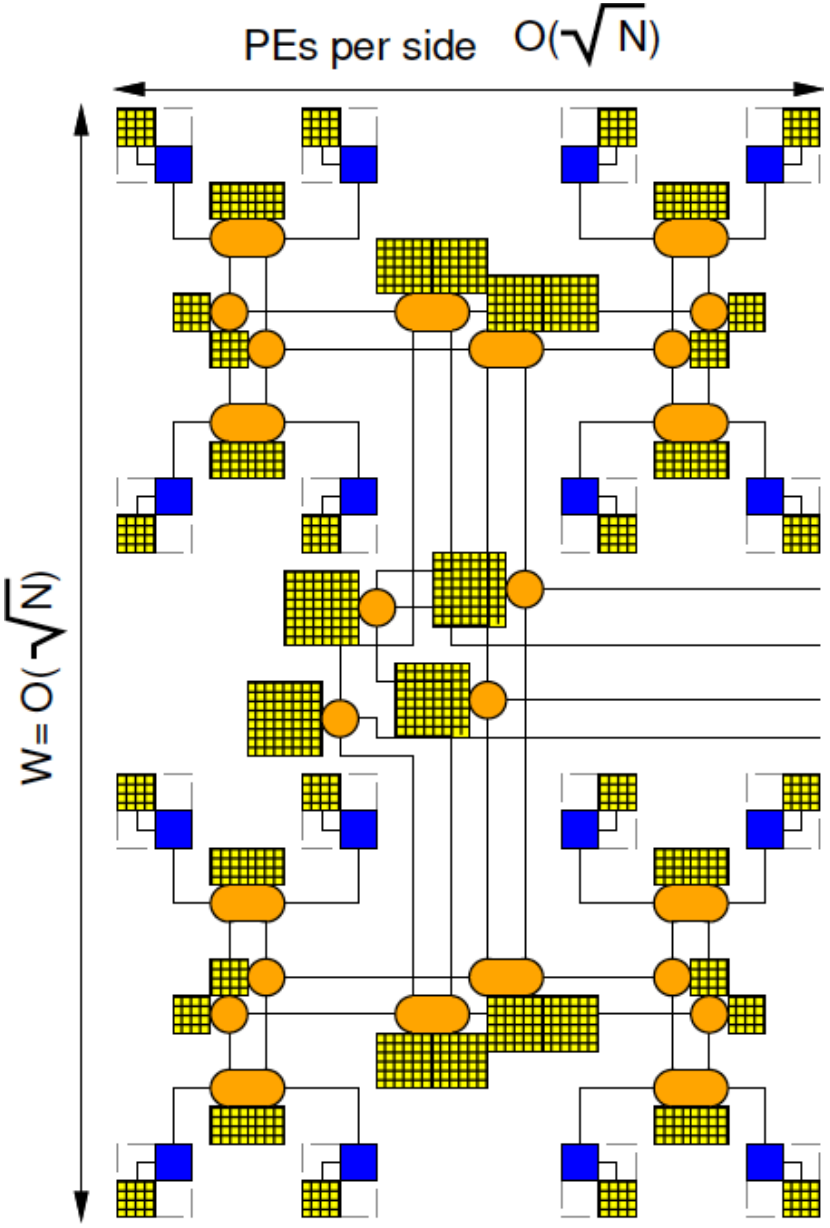
- Like an FPGA
- Each signal gets own wire
- No addresses
- Configuration local
- Area grows as $O(N^{2p})$ for $p>0.5$
- Energy $O(N^{2p})$ for $p>0.5$
  - $\Theta(N)$ for $p<0.5$
- Multilayer metal
  - Energy $O(N^{p+0.5})$ for $p>0.5$



(FPGA-like Architecture)

Instructions Local to Switches

- Constant metal
- Build p<0.5 tree
- Store bits local to each tree level
- Read out of memory there
- Bits/switch differs with tree level
- Signal on wire dominates reads
- $O(N^{p+0.5})$ for p>0.5



Results: Energy

Org	Any p	p<1.0	1>p>0.5	p=0.5	p<0.5
Processor	$O(N^{1.5}\log^{1.5}N)$	$O(N^{1.5}\log N)$ Description Locality			
Data Locality (Packet Switch)			$O(N^{p+0.5}\log^{1.5}N)$	$O(N\log^{2.5}N)$	$O(N\log^{1.5}N)$
FPGA 2-metal			$O(N^{2p})$	$O(N\log^{2}N)$	$\Theta(N)$
FPGA multilevel			$O(N^{p+0.5})$	$O(N\log N)$	$\Theta(N)$
Multicontext			$O(N^{p+0.5})$	$O(N\log N)$	$\Theta(N)$

Location^1, Location^2, Location^3

- To minimize asymptotic energy, essential to exploit spatial locality to:
  1. Reduce size of **description**
  2. Minimize **data** movement energy
     - Argues against centralized processor
  3. Reduce or eliminate **instruction** energy
     - Argues for configuration
       - Local to the resources controlled

Not Answered by This Paper

- Non-uniform activity
- Limited internal state
- SIMD word sharing of instructions
- Constants
  - Maybe O(1) sequential processors at leaves of tree OK? Better constants?
- Tighter lower bounds

Findings

- Oversimplifed soundbyte:
  - FPGAs use less energy than processors
- More precisely:
  - There is an asymptotic advantage to
    - Parallel, spatial evaluation
    - Over sequential evaluation on central processor
  - When
    - There is any spatial locality (Rent $p<1.0$)
    - There is uniform average gate activity
- With "enough" spatial locality ($p<0.5$), FPGAs use $\Theta(N)$ energy