ESE532: System-on-a-Chip Architecture

Day 23: April 12, 2017
Parallelism and Energy

Today

- How does parallelism impact energy?
- Refine
  - Locality?
  - Problem size?

Message

- Can tune parallelism to minimize energy
- Typically, the more parallel implementation costs less energy

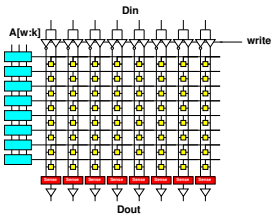
Day 22

Energy

- Growing domain of portables
  - Less energy/op → longer battery life
- Global Energy Crisis
- Power-envelope at key limit
  - E reduce → increase compute in P-envelope
  - Scaling
    - Power density **not** transistors limit sustained ops/s
  - Server rooms
    - Cost-of-ownership **not** dominated by Silicon
      - Cooling, Power bill

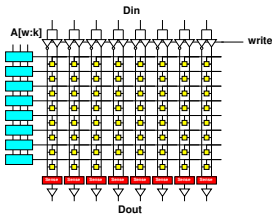
Memory Energy

- Memory operations cost energy
- Must move data for memory bit to outside of array
- Wires of length Sqrt(N)
- Energy Sqrt(N)



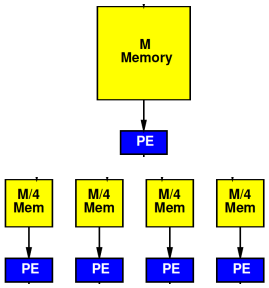
Memory Energy

- Energy Sqrt(N)
- Large memories cost more energy than small memories



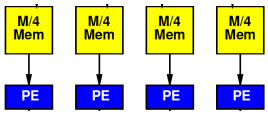
Preclass 1

- Energy
  - Per read from M=10^6 memory?
  - Per read from 10^6/4 memory?



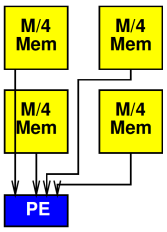
Local Consumption

- To exploit, we must consume the data local to the memory.



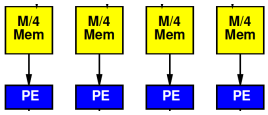
Cheat?

- What if we broke the memory into 4 blocks, but still routed to a single processor?



Exploit Locality

- Must consume data near computation

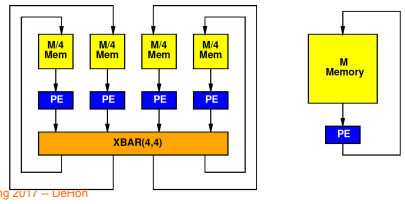


Inter PE Communication

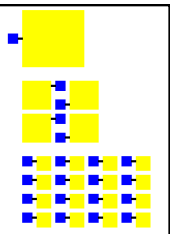
- May need to communicate between parallel processing units (and memories)
- Must pay for energy to move data between PEs

Preclass 2

- Energy: Read 4 memories 10^6/4, route 4x4 crossbar, write 4 10^6/4 memories?
- Energy: 4 reads from 10^6 memory, 4 writes from 10^6 memory?



Parallel Larger



- More parallel design
  - Has more PEs
  - Adds interconnect

→Total area > less parallel design

- More area → longer wires → more energy in communication between PEs
- Could increase energy!

Continuum Question

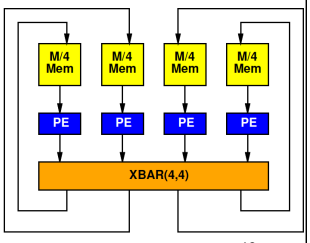
- Where do we minimize total energy?
  - Both memory and communication
- Design axis P – number of PEs
  - What P minimizes energy?

Simple Model

- E_{mem}=sqrt(M)
- Communication = E_{xbar}(I,O)=4*I*O
- P Processors
- N total data
- Possibly communicate each result to other PEs

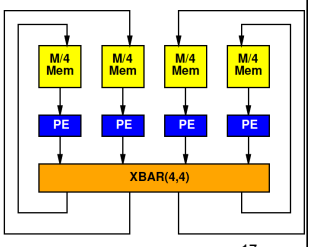
Simple Model: Memory

- Divide N data over P memories
- E_{mem}=sqrt(N/P)
- N total memory operations
- Memory energy: N*sqrt(N/P)
- Memory energy decrease with P



Simple Model: Communication

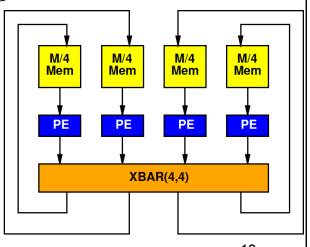
- Crossbar with P inputs and P outputs
- E_{xbar}=4*P*P
- Crossbar used N/P times
- Crossbar energy: 4*N*P
- Communication energy increase with P



Simple Model

$$N \times \sqrt{\left(\frac{N}{P}\right)} + N \times 4 \times P$$

$$N \times \left(\sqrt{\left(\frac{N}{P}\right)} + 4 \times P\right)$$



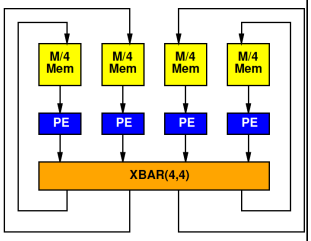
Preclass 3

- For N=10^6

$$N \times \left(\sqrt{\left(\frac{N}{P}\right)} + 4 \times P \right)$$

- Per operation becomes:

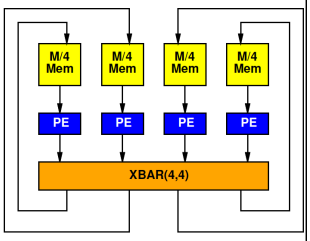
$$\left(\frac{10^3}{\sqrt{P}}\right) + 4 \times P$$



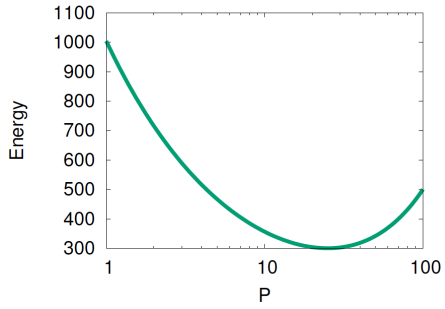
Preclass 3

- Energy for:
  - P=1
  - P=4
  - P=100
- Energy minimizing P?
  - Energy?

$$\left(\frac{10^3}{\sqrt{P}}\right) + 4 \times P$$



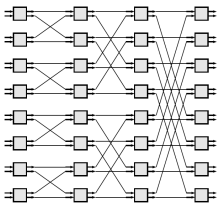
Graph

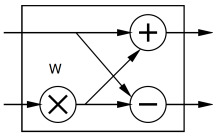


Day 14

FFT

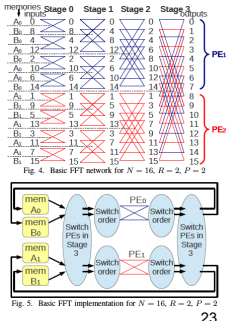
- Large space of FFTs
- Radix-2 FFT Butterfly





FFT Example

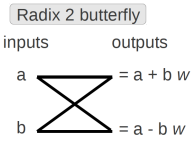
- Tune number of PEs, P



[Kadric, FCCM 2014]

FFT on Stratix-IV

- PE is Radix 2 Butterfly



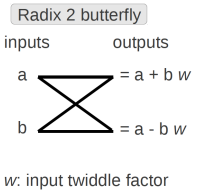
w: input twiddle factor

PEs	Hier	Mem	Components (per point)				Total
		Type	Mem	route	comp	route	pJ/pt
1	1	M144K	2300	140	900	1300	4700
1	16	M144K	1300	210	920	1400	3800
4	4	M9K	880	75	1000	1500	3500
16	1	M9K	420	85	900	1200	2600
32	1	M9K	210	85	910	1200	2400

[Kadric, FCCM 2014]

FFT

- PE is Radix 2 Butterfly

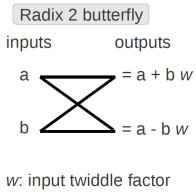


PEs	Hier	Mem	Components (per point)				Total
		Type	Mem	route	comp	route	pJ/pt
1	1	M144K	2300	140	900	1300	4700
1	16	M144K	1300	210	920	1400	3800
4	4	M9K	880	75	1000	1500	3500
16	1	M9K	420	85	900	1200	2600
32	1	M9K	210	85	910	1200	2400

[Kadric, FCCM 2014]

FFT

- PE is Radix 2 Butterfly

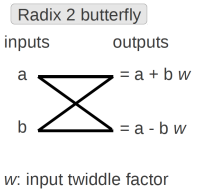


PEs	Hier	Mem	Components (per point)				Total
		Type	Mem	route	comp	route	pJ/pt
1	1	M144K	2300	140	900	1300	4700
1	16	M144K	1300	210	920	1400	3800
4	4	M9K	880	75	1000	1500	3500
16	1	M9K	420	85	900	1200	2600
32	1	M9K	210	85	910	1200	2400

[Kadric, FCCM 2014]

FFT

- PE is Radix 2 Butterfly

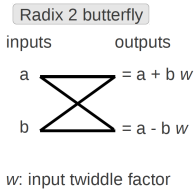


PEs	Hier	Mem	Components (per point)				Total
		Type	Mem	route	comp	route	pJ/pt
1	1	M144K	2300	140	900	1300	4700
1	16	M144K	1300	210	920	1400	3800
4	4	M9K	880	75	1000	1500	3500
16	1	M9K	420	85	900	1200	2600
32	1	M9K	210	85	910	1200	2400

[Kadric, FCCM 2014]

FFT

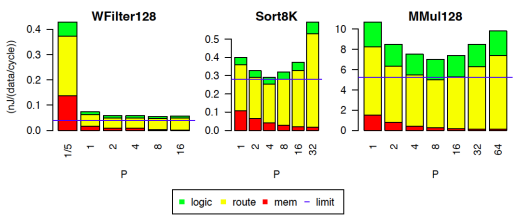
- PE is Radix 2 Butterfly



PEs	Hier	Mem	Components (per point)				Total
		Type	Mem	route	comp	route	pJ/pt
1	1	M144K	2300	140	900	1300	4700
1	16	M144K	1300	210	920	1400	3800
4	4	M9K	880	75	1000	1500	3500
16	1	M9K	420	85	900	1200	2600
32	1	M9K	210	85	910	1200	2400

[Kadric, FCCM 2014]

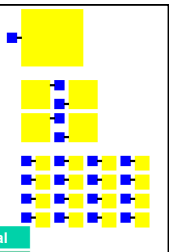
Tune Parallelism: Stratix-IV



[Kadric, FPGA 2015]

GMM

- Compute Gaussian model of pixel background for **every** pixel



PEs	Mem	Components (per PE)				Total
	Type	Mem	route	Comp	route	mW/PE
1	M144K	29	4.4	9.8	28	71
2	M144K	27	6.8	9.5	16	59
4	M9K	12	1.3	9.9	19	42
8	M9K	4.7	1.0	10	17	33
16	M9K	4.6	1.0	9.8	13	29
32	M9K	5.7	3.7	9.3	8	27

[Kadric, FCCM 2014]

GMM

- Compute Gaussian model of pixel background for **every** pixel

PEs	Mem	Components (per PE)				Total
	Type	Mem	route	Comp	route	mW/PE
1	M144K	29	4.4	9.8	28	71
2	M144K	27	6.8	9.5	16	59
4	M9K	12	1.3	9.9	19	42
8	M9K	4.7	1.0	10	17	33
16	M9K	4.6	1.0	9.8	13	29
32	M9K	5.7	3.7	9.3	8	27

[Kadric, FCCM 2014]

GMM

- Compute Gaussian model of pixel background for **every** pixel

PEs	Mem	Components (per PE)				Total
	Type	Mem	route	Comp	route	mW/PE
1	M144K	29	4.4	9.8	28	71
2	M144K	27	6.8	9.5	16	59
4	M9K	12	1.3	9.9	19	42
8	M9K	4.7	1.0	10	17	33
16	M9K	4.6	1.0	9.8	13	29
32	M9K	5.7	3.7	9.3	8	27

[Kadric, FCCM 2014]

GMM

- Compute Gaussian model of pixel background for **every** pixel

PEs	Mem	Components (per PE)				Total
	Type	Mem	route	Comp	route	mW/PE
1	M144K	29	4.4	9.8	28	71
2	M144K	27	6.8	9.5	16	59
4	M9K	12	1.3	9.9	19	42
8	M9K	4.7	1.0	10	17	33
16	M9K	4.6	1.0	9.8	13	29
32	M9K	5.7	3.7	9.3	8	27

[Kadric, FCCM 2014]

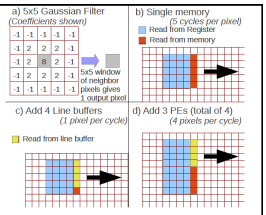
GMM

- Compute Gaussian model of pixel background for **every** pixel

PEs	Mem	Components (per PE)				Total
	Type	Mem	route	Comp	route	mW/PE
1	M144K	29	4.4	9.8	28	71
2	M144K	27	6.8	9.5	16	59
4	M9K	12	1.3	9.9	19	42
8	M9K	4.7	1.0	10	17	33
16	M9K	4.6	1.0	9.8	13	29
32	M9K	5.7	3.7	9.3	8	27

[Kadric, FCCM 2014]

Window Filter

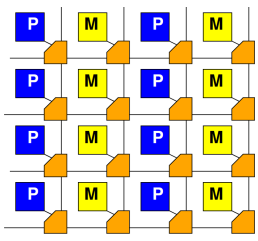


PEs	Line	Bits	Mem	Components (per pixel)				Total
	Buff	/pixel	Type	Mem	route	comp	route	pJ/PE
1	N	1	M144K	400	50	130	280	850
1	Y	1	M9K	110	11	81	120	320
2	Y	1	M9K	45	6	67	84	200
4	Y	1	M9K	24	3	61	75	160
4	Y	2	M9K	15	3	59	63	140
4	Y	4	M9K	15	3	57	60	135

[Kadric, FCCM 2014]

High Locality

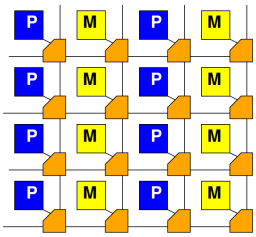
- If communication is local, don't need crossbar
- Communication energy scales less than P^2
- Can scale as low as P
- As see GMM, WinF



Model for High Locality

- E_{comm}=constant
- E_{comm}=10
- Total comm:
  N*10

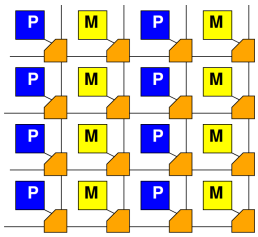
$$N \times \left(\sqrt{\left(\frac{N}{P} \right)} + 10 \right)$$



Preclass 4

- What is energy minimizing P?

$$\left(\frac{10^3}{\sqrt{P}} \right) + 10$$



Task Locality Matters

- Optimal P depends on communication locality
  - Very local problems always benefit from parallelism
  - Highly interconnected problems must balance energies → intermediate parallelism

Impact of Problem Size

- Optimal P changes with problem size
- Changes N
  - Changes cost of memories

$$N \times \left(\sqrt{\left(\frac{N}{P} \right)} + 4 \times P \right)$$

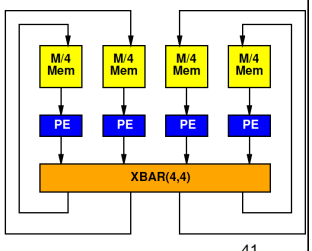
Preclass 5

- For N=6.4×10^7

$$N \times \left(\sqrt{\left(\frac{N}{P} \right)} + 4 \times P \right)$$

- Per operation becomes:

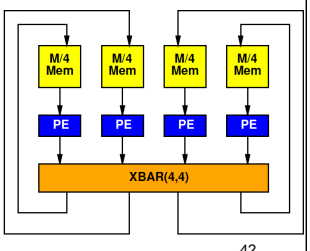
$$\left(\frac{8 \times 10^3}{\sqrt{P}} \right) + 4 \times P$$



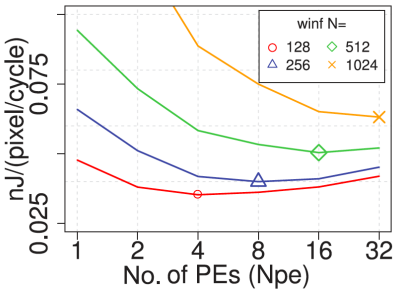
Preclass 5

- Energy minimizing P?

$$\left(\frac{8 \times 10^3}{\sqrt{P}} \right) + 4 \times P$$

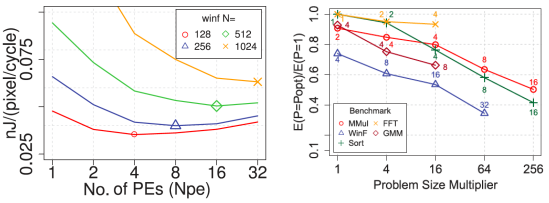


Window Filter



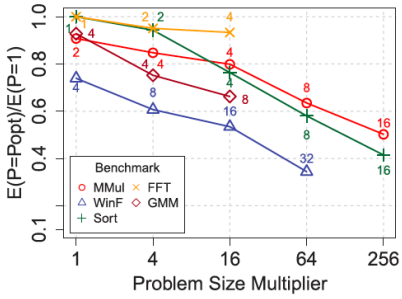
[Kadric, TRETS 2016]

PE Scaling with Problem Size



[Kadric, TRETS 2016]

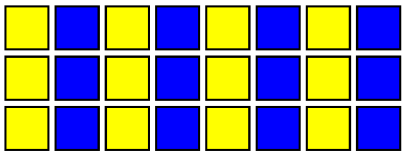
P Scaling Trends



[Kadric, TRETS 2016]

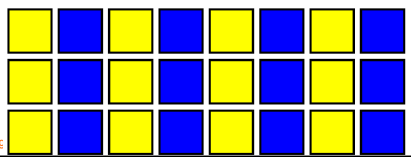
Balanced Memory and Compute

- In general
  - Want memory to be about the size of the PE
  - Then replicate that to handle larger problems
    - Rather than making memory larger



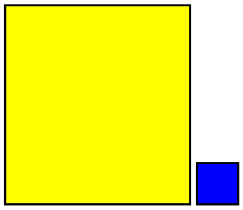
Balanced Memory and Compute

- Memory ~= Compute size
  - E_{mem} comparable to E_{route}
    - Doesn't make local computation much larger
  - If does need distant memory
    - Not much worse than one large memory



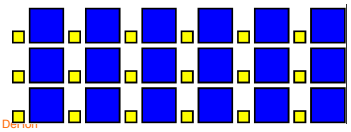
Balanced Memory and Compute

- Memory >> PE size
  - Not exploiting locality



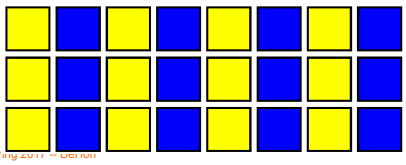
Balanced Memory and Compute

- Memory >> PE size
  - Not exploiting locality
- Memory << PE size
  - Energy routing to distant memory can be much larger than if each memory bank was larger



Balanced Memory and Compute

- In general
  - Want memory to be about the size of the PE
  - Then replicate that to handle larger problems
    - Rather than making memory larger
    - P α Problem Size



Big Ideas

- Large memories are energy expensive
  - Force large data movement
- Small memories with local compute
  - Reduce data movement, save energy
- Parallel design exploiting locality can reduce energy
- Optimal parallelism for problem
  - Driven by communication structure
  - Depends on problem size

Admin

- Project energy Milestone
  - Due Friday