Announcements:

- HW3
- Project
- Context

Application in Computation

→ P_{peak}: tasks/sec
I: tasks/customer
b: customers/sec

Translate the bank analogy to computers:

Which parts of this are task-dependent?

- P_{peak}
- b
- I

$$P \leq \min(P_{\text{peak}}, I \cdot b)$$

Hager et al. '17

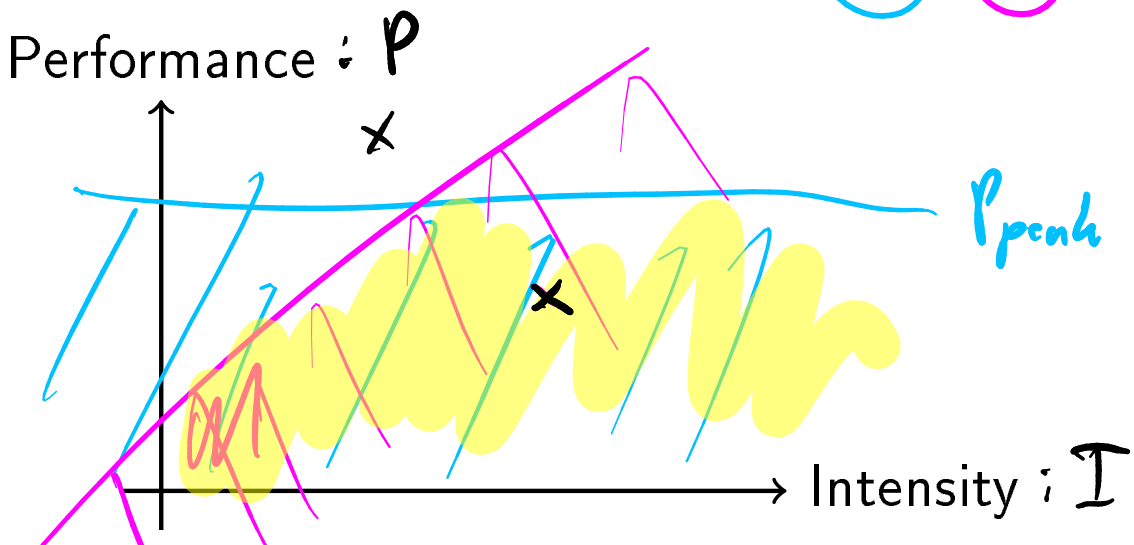
A Graphical Representation: 'Roofline'

Plot (often log-log, but not necessarily):

- X-Axis: Intensity
- Y-Axis: Performance

What does our inequality correspond to graphically?

$$P \leq \min(P_{\text{peak}}, I \cdot b)$$



What does the shaded area mean?

"In": attainable

"Out": not attainable

Example: Vector Addition

```
float
double r, s, a[N];
for (i=0; i<N; ++i)
  a[i] = r + s * a[i];}
```

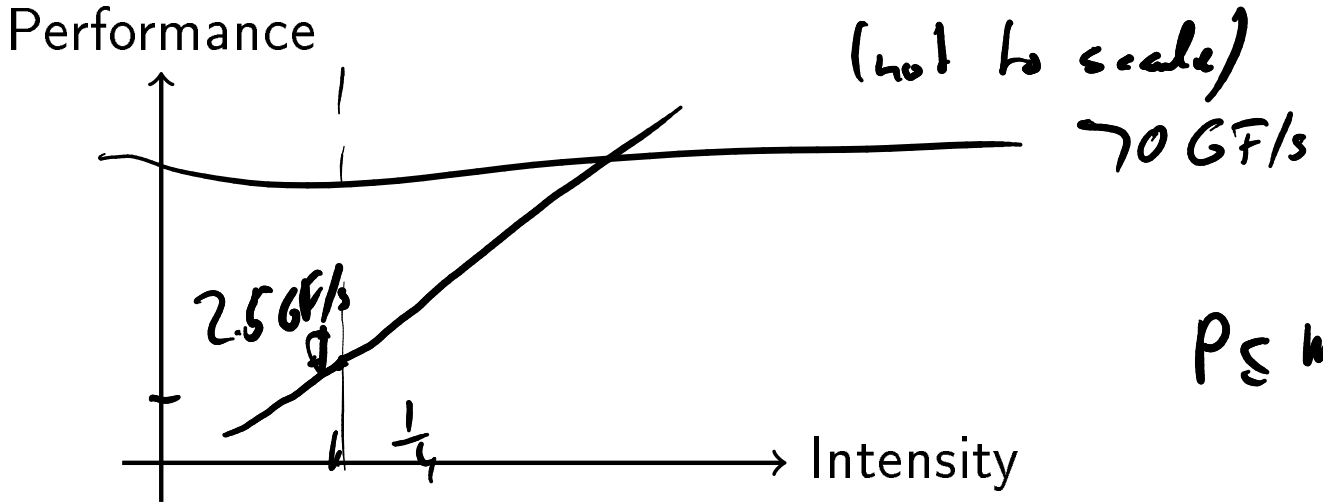
Find the parameters and make a prediction.

M

$b \approx 10$ GB/s (? dep on accounting for writes)

$P_{peak} \approx 70$ GF/s

$I = \frac{2 \text{ flops}}{8 \text{ bytes}}$ $\frac{F}{B} = \frac{1}{4} \frac{F}{B}$



$P \le \min(I \cdot b, P_{peak})$

Refining the Model

- P_{max}: Applicable peak performance of a loop, assuming that data comes from the fastest data path (this is not necessarily P_{peak})
- Computational intensity ("work" per byte transferred) over the slowest data path utilized
- b: Applicable peak bandwidth of the slowest data path utilized

Hager et al. '17

Calibrating the Model: Bandwidth

Typically done with the STREAM benchmark.

Four parts: Copy, Scale, Add, Triad `a[i] = b[i] + s\cdot c[i]`

Do the four measurements matter?

no, weak attempt at modelity I

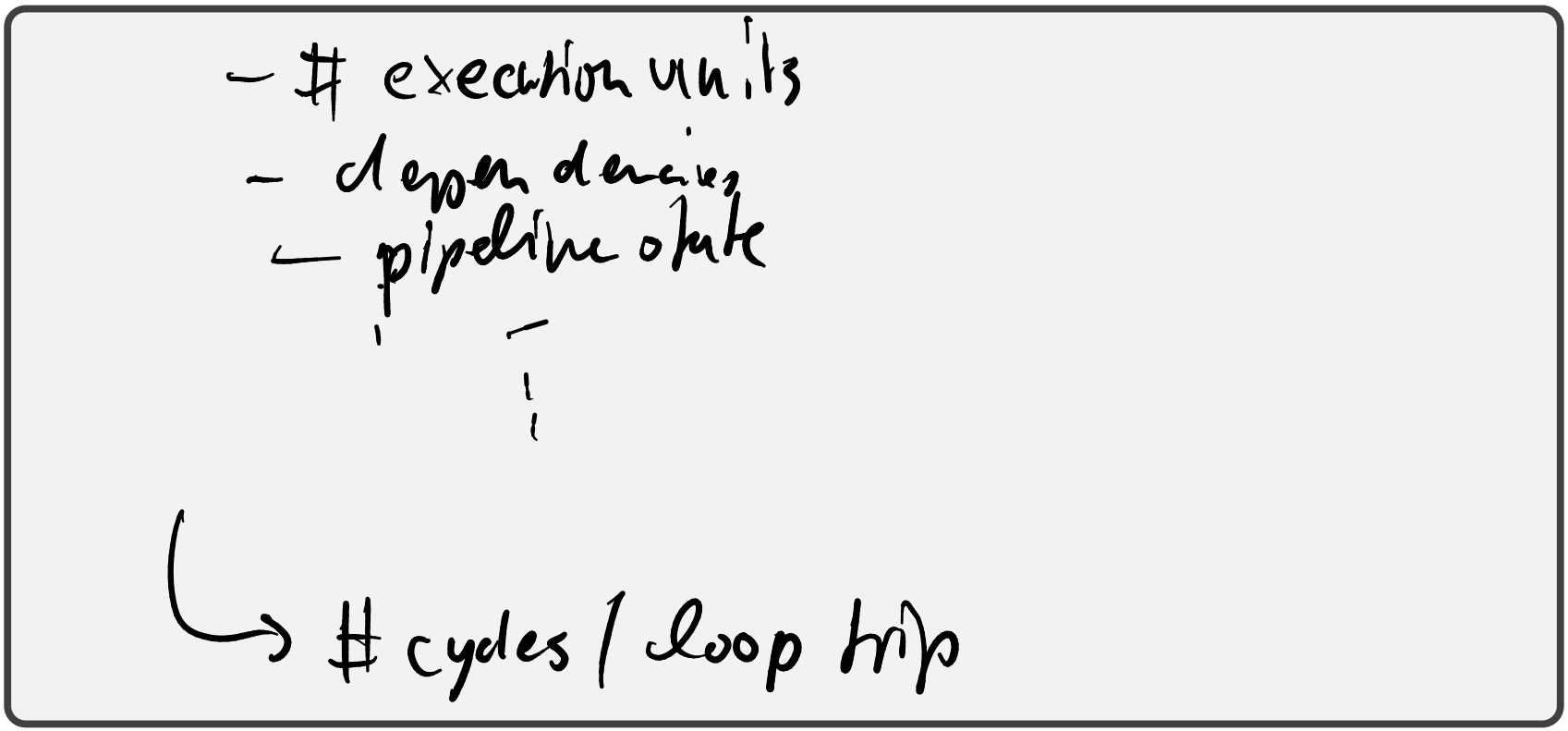
Any pitfalls?

non-temporal stores

McCalpin: STREAM

Calibrating the Model: Peak Throughput

Name aspects that should/could be factored in when determining peak performance:



Practical Tool: IACA

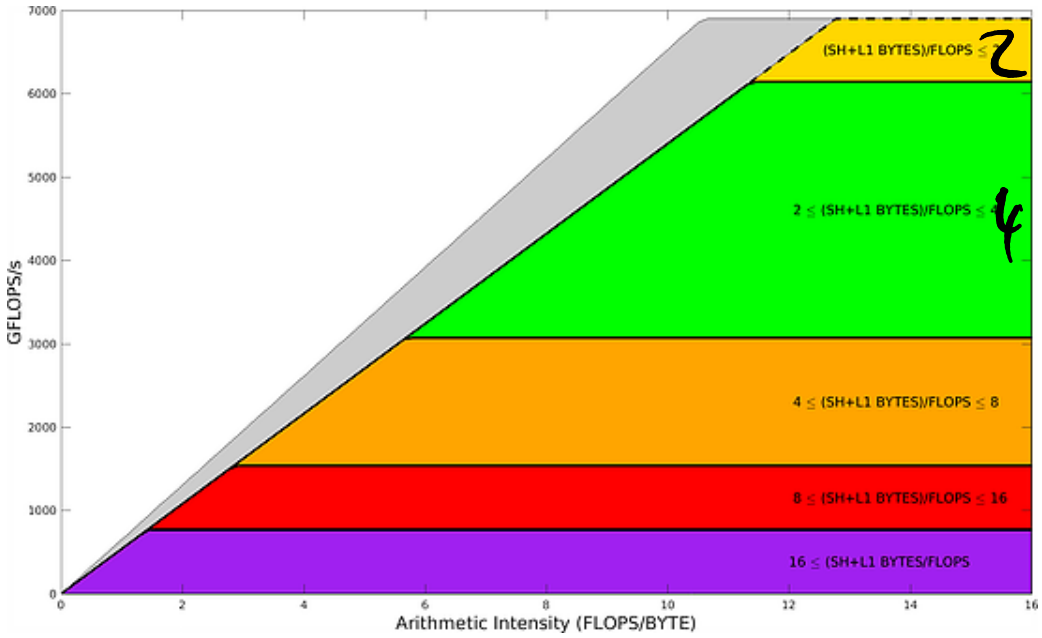
Question: Where to obtain an estimate of P_{max}?

Demo: perf/Forming Architectural Performance Expectations

Questions:

- What does IACA do about memory access? / the memory hierarchy?

An Example: Exploring Titan V Limits

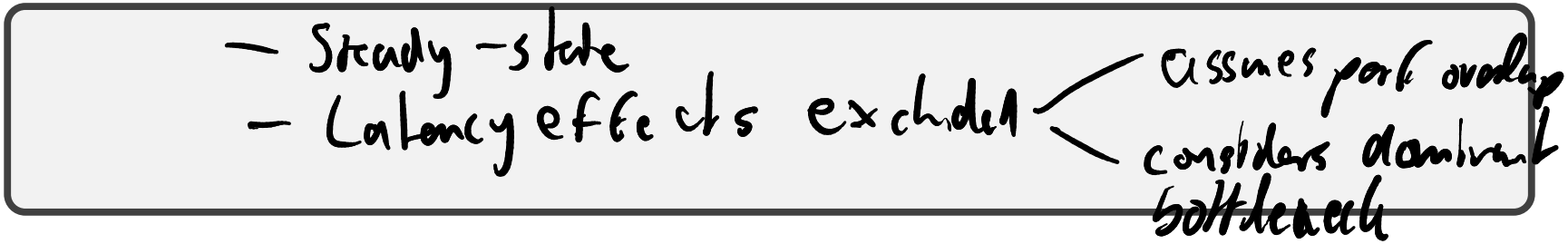


- Memory bandwidth: 652 GB/s theoretical, 540 GB/s achievable
- Scratchpad / L1 throughput:
  80 (cores) x 32 (simd width) x 4 (word bytes) x 1.2 (base clock) ~= 12.288 TB/s
- Theoretical ~~peak~~ flops of 6.9 TFLOPS/s [Wikipedia]

Warburton '18

Rooflines: Assumptions

What assumptions are built into the roofline model?



Important to remember:

- It is what you make of it–the better your calibration, the more info you get
- But: Calibrating on experimental data loses predictive power (e.g. SPMV)

Modeling Parallel Speedup: A 'Universal' Scaling Law

Develop a model of *throughput* $X(N)$ for a given load N, assuming execution resources scale with N.

$$X(N) = \frac{\gamma \cdot N}{1 + \alpha \cdot (N-1) + \beta N(N-1)}$$

α: models queuing

β: models incoherence (all-to-all)

γ: "perfect" scaling

[Gunther '93]

Outline

Timing Experiments: Pitfalls

What are potential issues in timing experiments? (What can you do about them?)

- Compiler opt.
- Overheads
- Timing noise
  - Other people
  - timer resolution
  - "chunk size" sufficiently big
  - know your clock sources
    RTC, monotonic, process, user,
    TSC (nominal, core clock)

Timing Experiments: Pitfalls (part 2)

What are potential issues in timing experiments? (What can you do about them?)

Combining Multiple Measurements

How can one combine multiple performance measurements? (e.g. "average speedup"?)

Example: Which computer should you buy?

Execution time [s]	Computer A	Computer B	Computer C
Program 1	1	10	20
Program 2	1000	100	20