# Assignment 8 – Task 1: Resource-driven modeling

Loop:

- $N$ iterations (it.)
- $V$ bytes/it. of data traffic
- $W$ instr./it. of processor work
- $b_S$ memory bandwidth in byte/s
- $p_E$ instruction throughput in instr./cy
- $f$ clock frequency

- Execution time: $T_{exec} = \frac{W}{p_E}$ [cy/it.]
- Data transfer time: $T_{data} = \frac{V \times f}{b_S}$ [cy/it.]

# Assignment 8 – Task 1: Resource-driven modeling

Predictions in cy/it.:

a) Non-overlapping: $T_{NO} = T_{exec} + T_{data} = \frac{W}{p_E} + \frac{V \times f}{b_S}$

b) Overlapping: $T_O = \max(T_{exec}, T_{data}) = \max\left(\frac{W}{p_E}, \frac{V \times f}{b_S}\right)$

Model for execution performance $P$ in instr./s:

c) $P = \frac{W \times f}{T}$, so

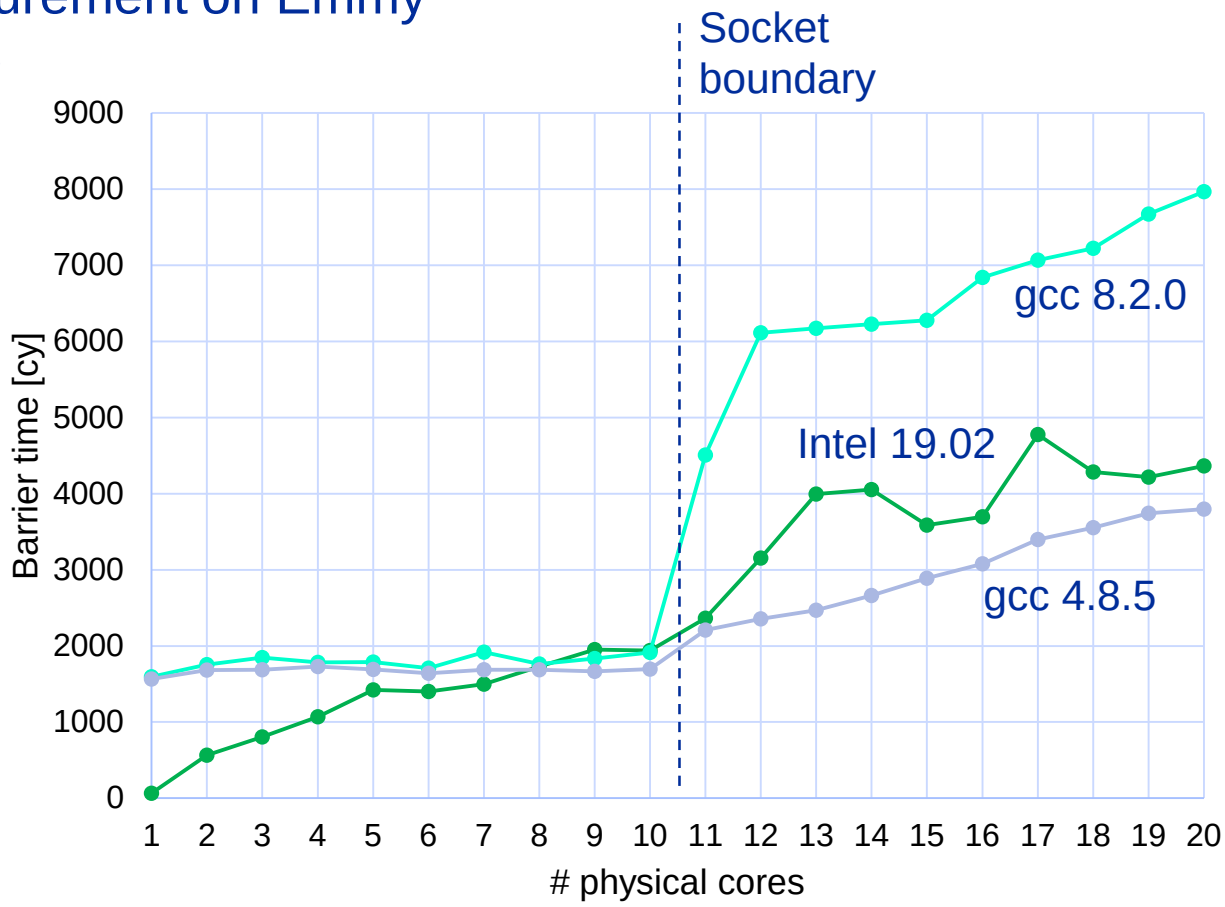
$$P_{NO} = \frac{W \times f}{W/p_E + V \times f/b_S} = \left(\frac{1}{p_E \times f} + \frac{V}{W \times b_S}\right)^{-1}$$

$$P_O = \frac{W \times f}{\max\left(\frac{W}{p_E}, V \times \frac{f}{b_S}\right)} = \min\left(p_E \times f, \frac{W}{V} \times b_S\right)$$

“intensity”

# Assignment 8 – Task 2

a) Barrier Measurement on Emmy (Intel vs. gcc)




Intel19, 40 threads (incl. SMT): 5500-6500 cy

# Assignment 8 – Task 2

b) Barrier overhead for

- IVB dual-socket (40 threads): 6000 cy
- Intel Xeon Phi (240 threads): 18000 cy

That does not look bad for Xeon Phi. Or does it???

3 x cores (20 vs 60) on Phi
2 x more operations per cycle per core on Phi (16 F/cy instead of 8 F/cy)

→ 6 x more work done on full Xeon Phi per cycle

3 x more barrier penalty (cycles) on Phi

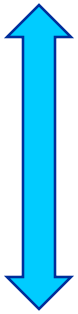
**→ One barrier causes 3 x 6 = 18 x more pain ☺.**

# Assignment 8 – Task 2

c) Performance prediction: Schönauer triad, N=500, 10 threads

```
A(:)=B(:)+C(:)*D(:)
```

Pure execution in L1 cache:
Assuming AVX, IVB core: 1 LD + ½ ST per cycle
→ 4 it / 3 cy ( = 5.83 Gflop/s @ 2.2 GHz)
→ 500 it. on 10 cores: 500 it / (10 * 4 it / 3 cy) = 37.5 cy



Including barrier overhead:
Barrier on 10 cores: 2000 cy
→ Overall time: T = 2037.5 cy
→ Performance:
P = 1000 flops / 2037.5 cy * 2.2 Gcy/s = 1.08 Gflop/s
(instead of 58.3 Gflop/s)

# Assignment 8 – Task 3: Roofline modeling

Starting point for Roofline Model: Code balance (inverse intensity)
Double precision, inner loop over index **`i`** implied!

a) **`a[i] = b[i] + c[i]`** (STREAM add)

$B_c$ = **24 B/F** (using NT stores), $P_{max,core}$ = 4F / 2cy (AVX)

b) **`s = s + a[i] / b[i]`** ("scalar ratio")

$B_c$ = **8 B/F,** $P_{max,core}$ = 8F / 28cy (AVX)

c) **`y[i] = y[i] + v[j][i] * b[k[i]]`** (JDS spMVM)

**14 B/F** < $B_c$ < **46 B/F**
If **`k[i]=i`** we have $B_c$ = **18 B/F,** $P_{max,core}$ = 2F / 2.5cy (scalar)

# Assignment 8 – Task 3: Roofline modeling

Roofline bandwidth parameter (10 cores): $b_S \approx 40$ GB/s

Clock speed: $f = 2.2$ GHz

a) STREAM ADD:

$$P = \min(10 \cdot f \cdot \frac{4}{2} {}^{F}/_{cy}, {}^{b_S}/_{B_C}) = \min(44, 1.66) \, {}^{GF}/_{s} = 1.66 \, {}^{GF}/_{s}$$

→ Memory bound!

BW for full „peak“: $b_S = B_c \cdot 10 \cdot f \cdot 2$ F/cy ≈ 1056 GB/s

# Assignment 8 – Task 3: Roofline modeling

b) Scalar ratio:
Peak performance is dominated by divide (28 cy / 8 F)

$$P = \min\left(10 \cdot f \cdot {}^{8}/_{28}\,{}^{F}/_{cy}\,, {}^{b_S}/_{B_C}\right) = \min(6.28, 5)\,{}^{GF}/_{s} = 5\,{}^{GF}/_{s}$$

→ Memory bound! (but just barely)

c) JDS spMVM:

$$P = \min\left(10 \cdot f \cdot \frac{4}{5}\,{}^{F}/_{cy}\,, {}^{b_S}/_{B_C}\right) = \min\left(17.6, \begin{cases} 2.86 \\ 2.22 \\ 0.87 \end{cases}\right) {}^{GF}/_{s}$$

→ Memory bound!

Actual performance may be worse due to issues with prefetching and TLB misses