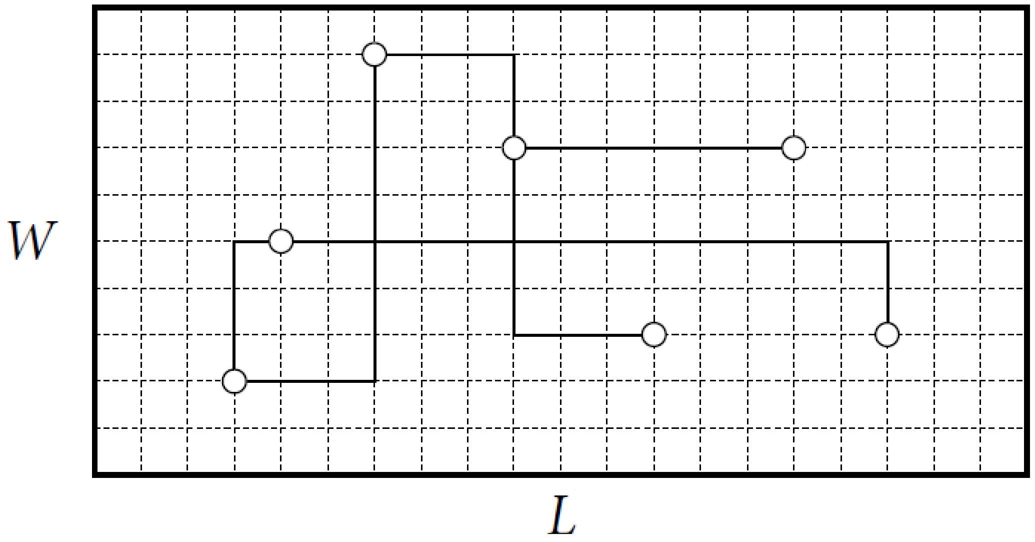
Comparing Classic VLSI Bounds for Matrix Multiplication with Loomis-Whitney Bound (just an exercise to aid understanding)

Jonathan Greene, 3/2/22, 11am

VLSI Complexity (Thompson, 1980)

- Computation network (circuit) for g is layed-out on grid:



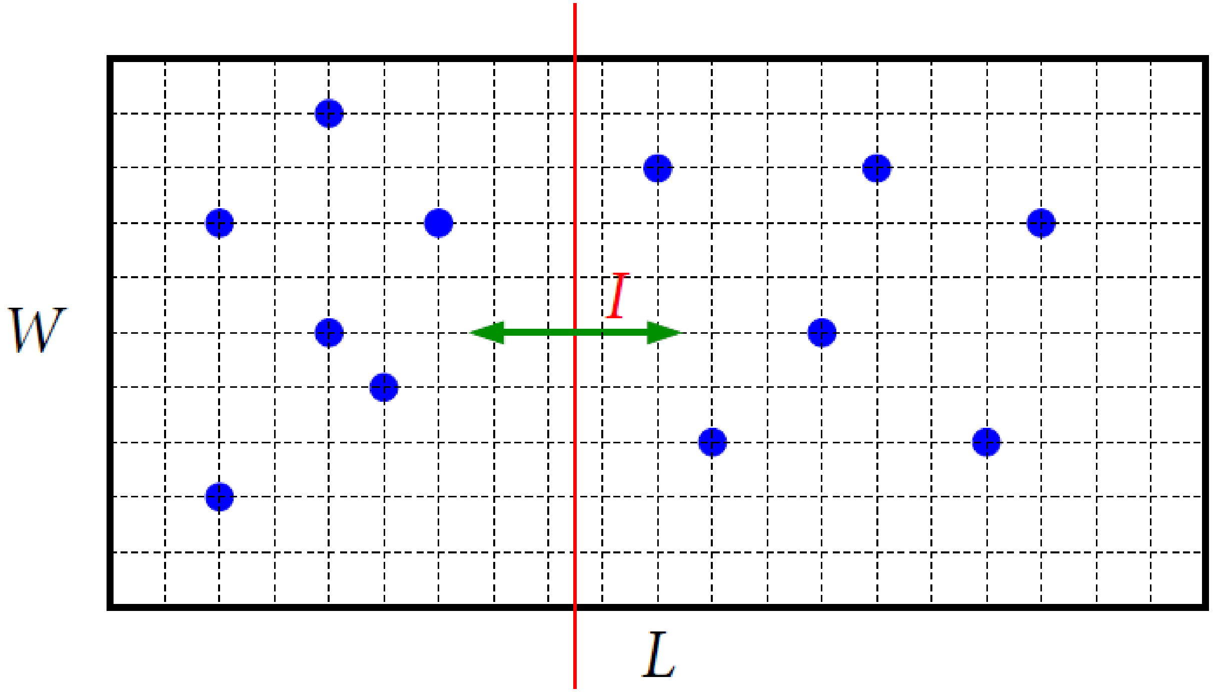
- Grid point: input, output, logic/memory element, wire crossing
- Grid line: constant number of wires
- Each wire carries constant number of bits/clock cycle
- Chip area for g: $A = W \times L$ grid squares $(W \leq L)$
- Computation time for g: T clock cycles

Essence of the model:

- Logic is free
- Communication takes time and area

Slides: A. El Gamal

- Thompson (1980) used cutset argument to obtain lower bound on AT^2

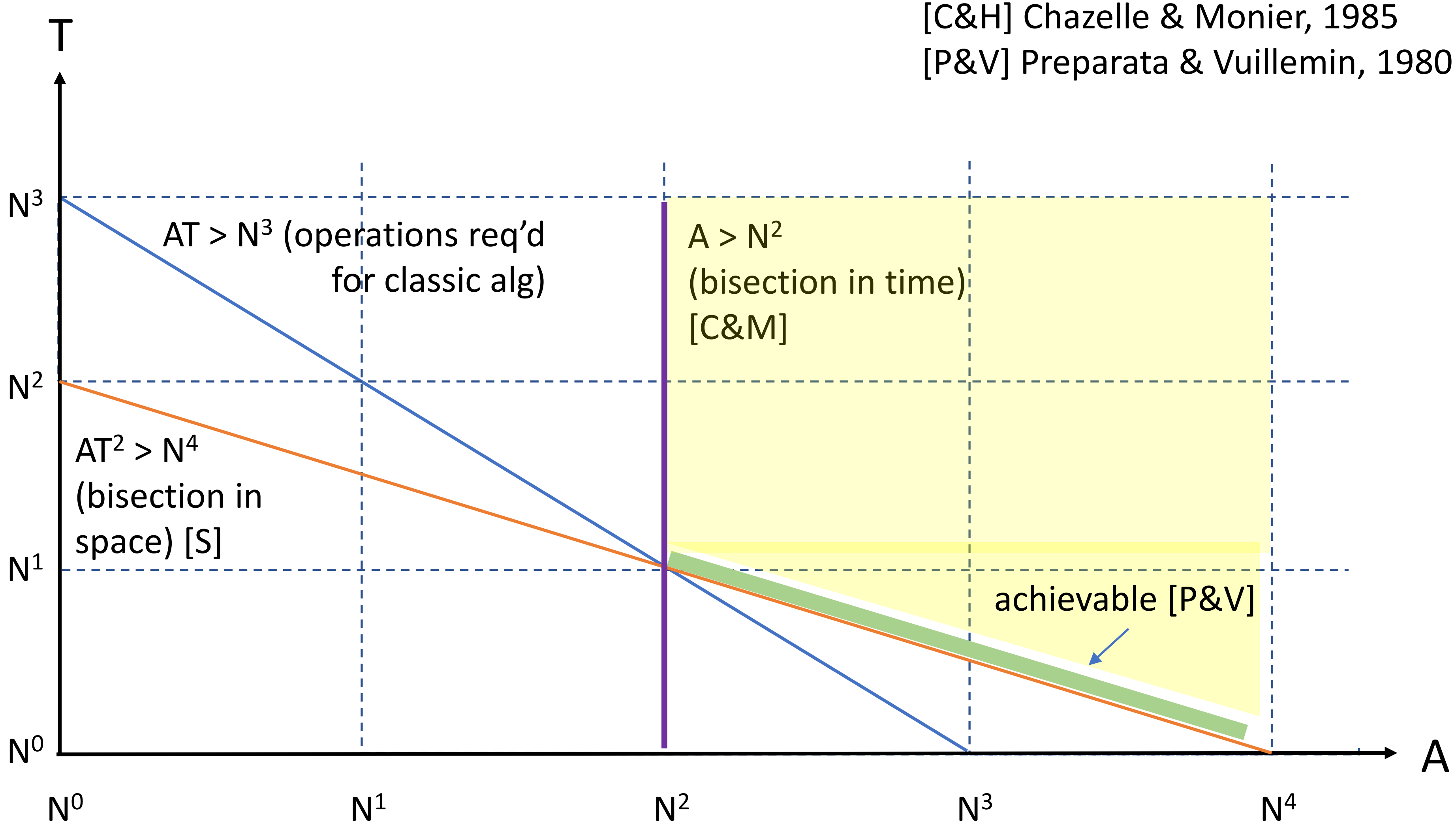


  - Bisect chip such that each side has $n/2$ inputs
  - Let I be min # of bits exchanged in any chip that computes g
- $I \leq cWT \leq c\sqrt{A}\,T \;\Rightarrow\; AT^2 = \Omega(I^2)$
- Tight for sorting, DFT, matrix multiplication, . . .

Can also derive bounds on power

VLSI Bounds for Classic NxN Matrix Multiplication

[S] Savage, 1981
[C&H] Chazelle & Monier, 1985
[P&V] Preparata & Vuillemin, 1980



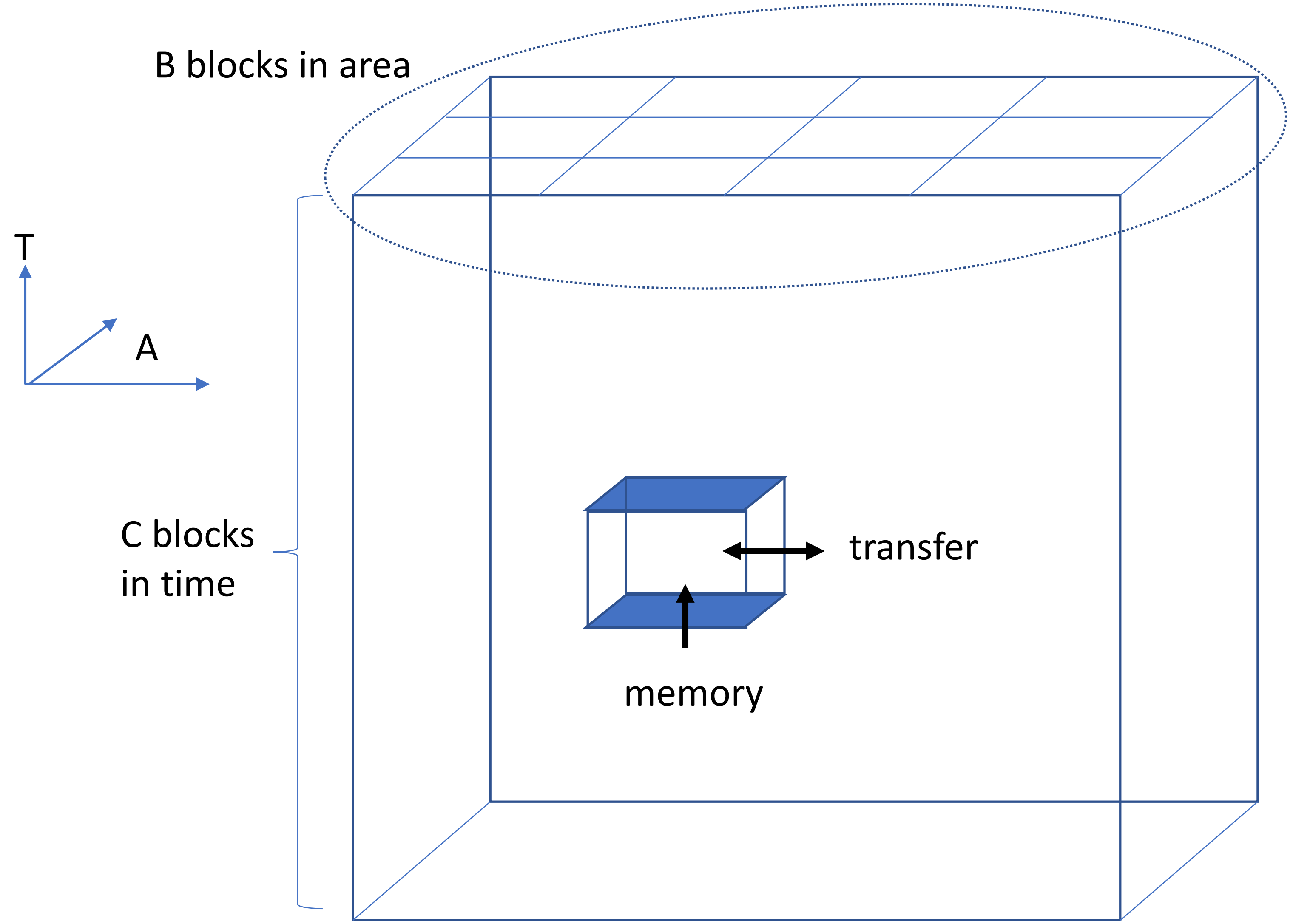
$AT > N^3$ (operations req'd for classic alg)

$A > N^2$ (bisection in time) [C&M]

$AT^2 > N^4$ (bisection in space) [S]

achievable [P&V]

Attempt to apply Loomis-Whitney in VLSI Context



Notation: “<“ refers to asymptotic growth rate

Total blocks = BC

Memory per block < A/B
(or 0 if C = 1)
Transfer per block < $(A^{\frac{1}{2}} / B^{\frac{1}{2}})\ (T/C)$
(or 0 if B = 1)
Ops per block < (A/B)(T/C)

There must exist a block with:

- Share of inputs < 1/(BC)
- Share of ops > 1/(BC)

For matrix multiplication:

Total inputs = N^2

Ops = N^3

To perform M ops, need $M^{2/3}$ word of data (Loomis & Whitney)

For any B and C, $1 < B < A$ and $1 < C < T$, there exists a block with:

Memory $< A\, B^{-1}$

Transfer $< A^{½}\, T\, B^{-½}\, C^{-1}$

Inputs $< N^2\, B^{-1}\, C^{-1}$

Ops $> N^3\, B^{-1}\, C^{-1}$

Memory + Transfer + Inputs > $(\text{Ops})^{2/3}$

$$A\, B^{-1} + A^{½}\, T\, B^{-½}\, C^{-1} + N^2\, B^{-1}\, C^{-1} > N^{2}\, B^{-2/3}\, C^{-2/3}$$

which implies

$$A > N^2\, B^{1/3}\, C^{-2/3} \quad \text{or} \quad AT^2 > N^4\, B^{-1/3}\, C^{2/3}$$

When B C = 2, this is similar to classic bounds.
Choosing any other values for B and C doesn’t improve things.

The Loomis-Whitney bound assumes a classic matrix multiply algorithm with N^3 operation.

The [C&M] and [S] bisection bounds seem just as strong in a VLSI context, but do not require that assumption.

Is it possible we can usefully apply those bounds in a processor/memory context?