Where are we now, and where to go?

Jun Makino
Exascale Computing Project
AICS, RIKEN

Five eras of evolution of CPUs

I —1969: Before CDC7600 (fully pipelined multiplier)

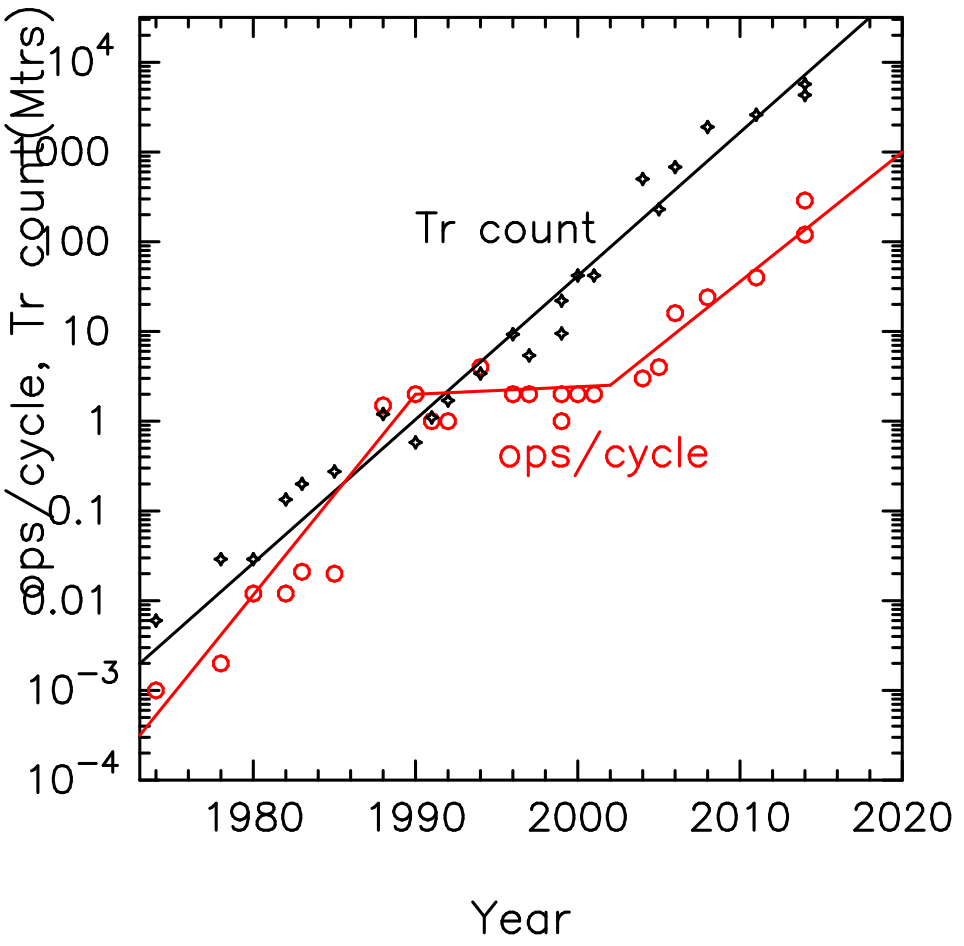
II —1989: Before Intel i860 (single-chip CPU with (almost) fully pipelined multiplier)

III —2003:CMOS scaling era (Power $\propto$ size^3) From i860 to Pentium 4

IV —2022(?):Post-CMOS scaling era (Power $\propto$ size) From Athlon 64 X2 to Knights Hill(?)

V 2022(?)—: Post-Moore era(miniaturization stops) ???

Evolution of Microprocessors



Black: # of transistors in CPU chips

Red: # of FPUs in CPU chips

Quick rise in era II, no increase in era III, and steady increase in era IV.

Observations

- **In era III (1989-2003), number of FPUs in CPU chips did not increase, even though the number of transistors in CPU chips increased by more than a factor of 100.**
- **In era IV (2003-2022?), the number of FPUs has been proportional to the number of transistors**

Observations

- **In era III (1989-2003), number of FPUs in CPU chips did not increase, even though the number of transistors in CPU chips increased by more than a factor of 100.**
- **In era IV (2003-2022?), the number of FPUs has been proportional to the number of transistors**

So what can we do in era V?

Not all processors are created equal

How many ~~Intel engineers~~ transistors it takes to do one floating-point op?

GRAPE-3	1991	4K	apl-specific pipeline
GRAPE-6	1997	30K	apl-specific pipeline
GRAPE-DR	2006	400K	SIMD 512-core chip
Cray-1	1976	400K	early vector
Intel i860	1989	600K	Beginning of era III
Earth Simulator	2002	4M	matured vector
NV Fermi	2010	3M	GPGPU
Sandy Bridge	2011	40M	Deep in era IV

Difference of 3-4 orders of magnitude

Why such a huge difference?

- Two orders of magnitude for programmable cores with double-precision arithmetic
- Two orders of magnitude from number format, specialized pipeline, etc

Possible sources of difference in programmable cores

- Deep pipeline
- Register file
- Cache
- Instruction fetch/decode/....
- Memory interface

So how many transistors do you really need to do one multiplication?

Mantissa	# trs.
53	$\sim$ 100K
23	$\sim$ 20K
16	$\sim$ 10K
8	$\sim$ 3K

Quite a large room for "improvement"

How about power consumption

- power $\propto$ # of transistors
- therefore reduction of transistor count per operation means reduction of power
- low-voltage operation (NTV etc) would give another order of magnitude improvement

Speculations

- There *might be some* life after 5nm, even without any exotic technology
- 1-2 orders of magnitude by streamlining the processor architecture
- 1-2 orders of magnitude by optimizing the word length (need to develop new algorithms)
- 1 order of magnitude by going to low voltage
- In total, 3-5 orders of magnitude = 15-25 years.
- Probably more than enough to cover my retirement and beyond