Poor Man's Computing: how much power you can get from a Linux PC

Alexander Shchepetkin

- Specifics of architecture: strong and weak points
- Software side of the story: availability, performance
- Specific issues of code optimization
- Comparison with SGI ORIGIN 2000, Sun Enterprise
- What to expect next

Computer #1: 2 $\times$ 933 MHz Pentium IIIs on ASUS CUV4X-D (VIA 694XP set), 1024MB PC-133 memory (1.06 GB/sec bandwidth);

- 256 KB 8-way (!) set-associative cache
- complex instruction set (not RISC)
- SMP-architecture (shared bus like first SGIs)
- **multi bootable:** Linux Mandrake 8.2 "OEM Standard" (running 2.4.18-8.1smp.mdk kernel); and Linux Mandrake 7.2 "Complete" (2.2.17smp kernel) operating systems; as well as Windows 2000.
- GNU `gcc`, `g77` compilers (F77 but not F90) **FREE**
- Intel `icc` and `ifc` compilers: F95 support; Open MP support; **FREE**
- Lahey (Fujitsu) `lf95` compiler, striped down version (no `OpenMP` support) $240; Complete version $640

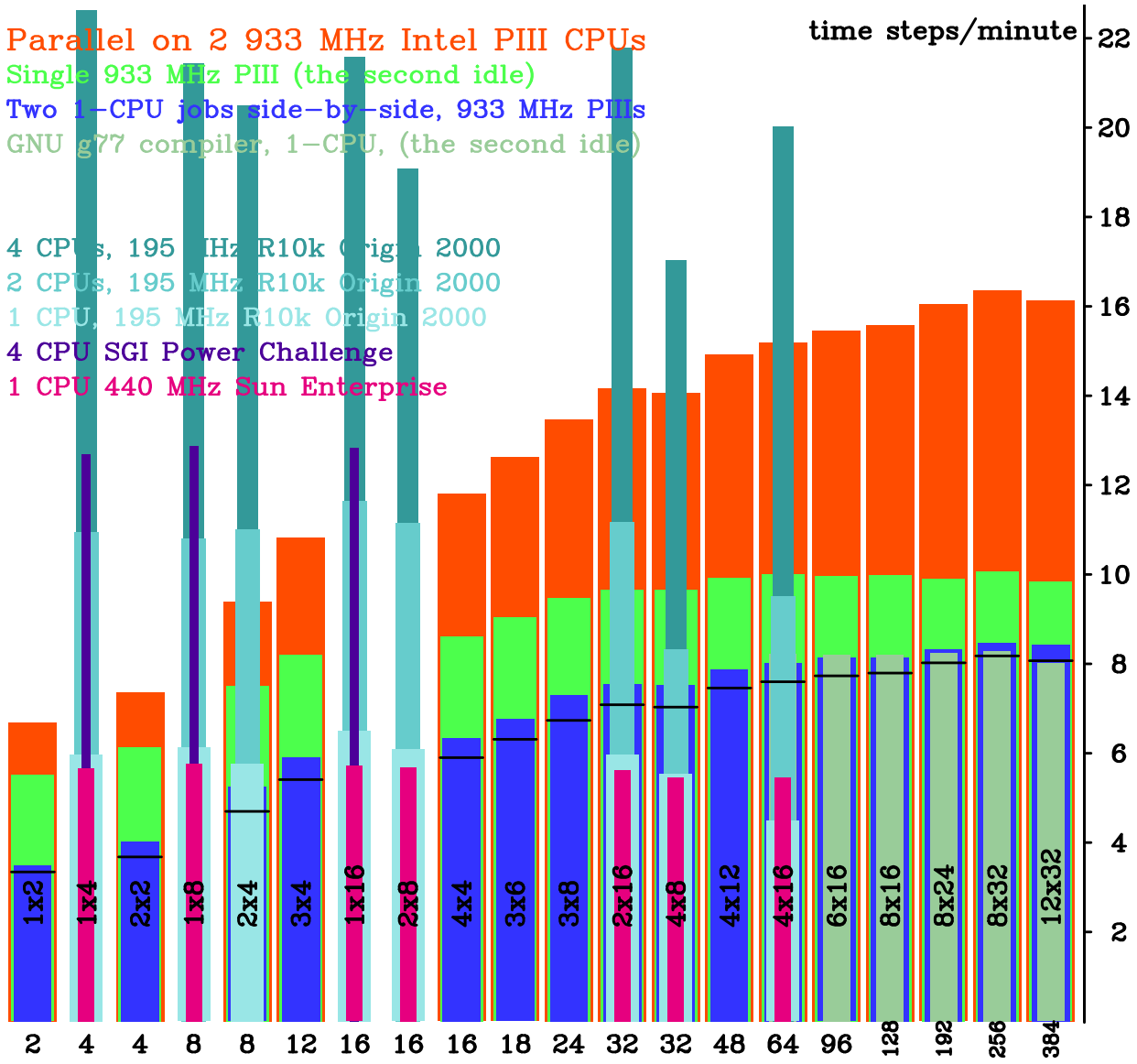
- State of the art in Jan 2001, not so impressive today

Computer #2: 2 $\times$ 2.4 GHz Intel Xeon CPUs on Supermicro P4DCE+ board (i860 chip set), 1024MB PC800 RDRAM memory (3.2 GB/sec bandwidth);

- 512 KB cache ;4x cache line relatively to PIII
- Linux Mandrake 9.1 (2.4.21-024smp/ent kernel)
- Intel `ifc/icc` 6.0.1-304 and 7.1.0xx compilers
- Cost to build $1700 (beginning 2003)

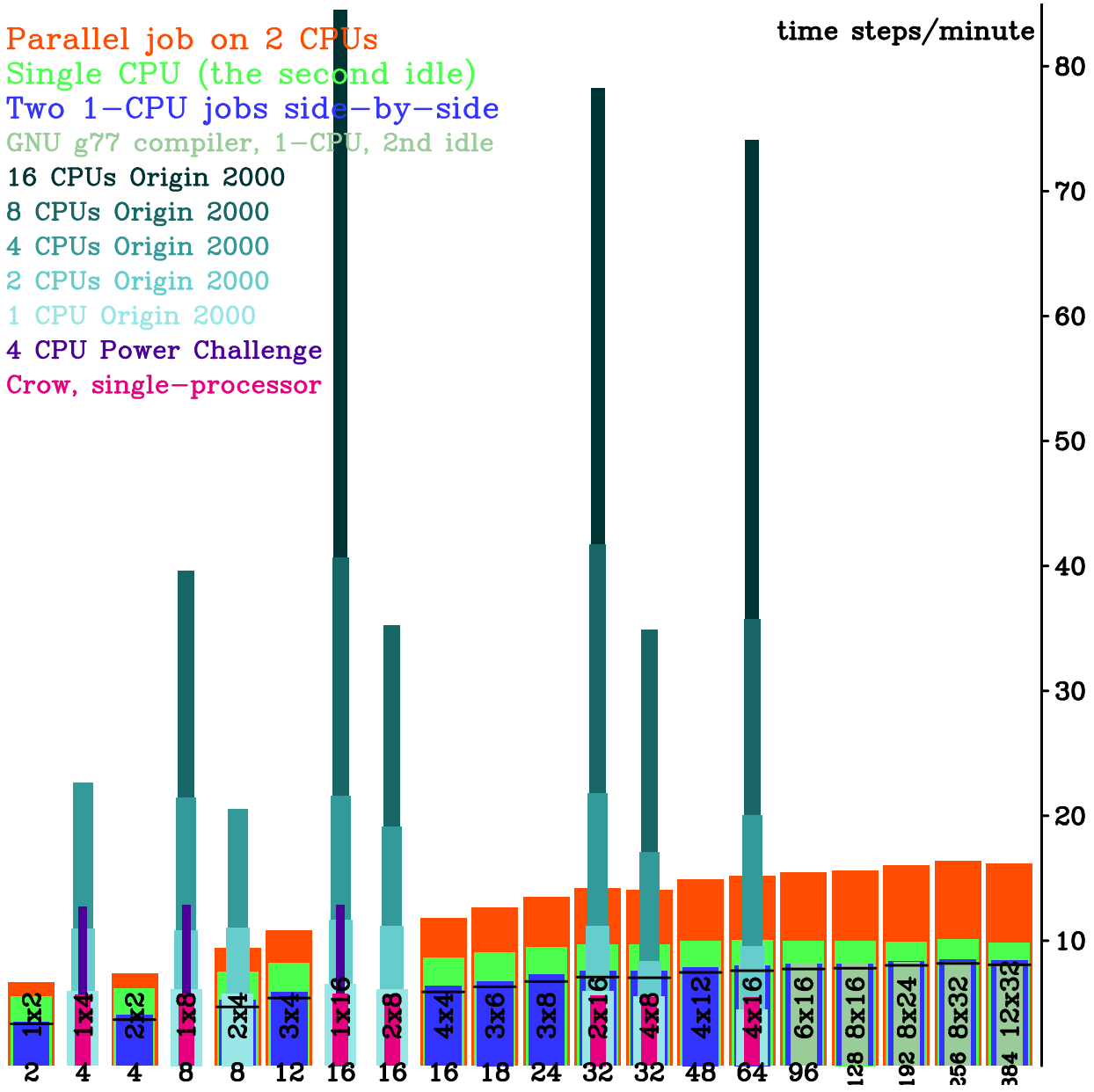
Test Problem:

- ROMS code:
  - subdomains decomposition capability
  - number of subdomains is independent from number of CPUs used
  - zig-zag tile processing order
  - all numerical features of ROMS
- 3/4 degree North - Equatorial Atlantic DAMEE configuration, 128 $\times$ 128 $\times$ 20 grid, $\Rightarrow$ 100MB problem
- 1/2 degree Pacific model, 384 $\times$ 224 $\times$ 30 grid, $\Rightarrow$ 800MB+ problem



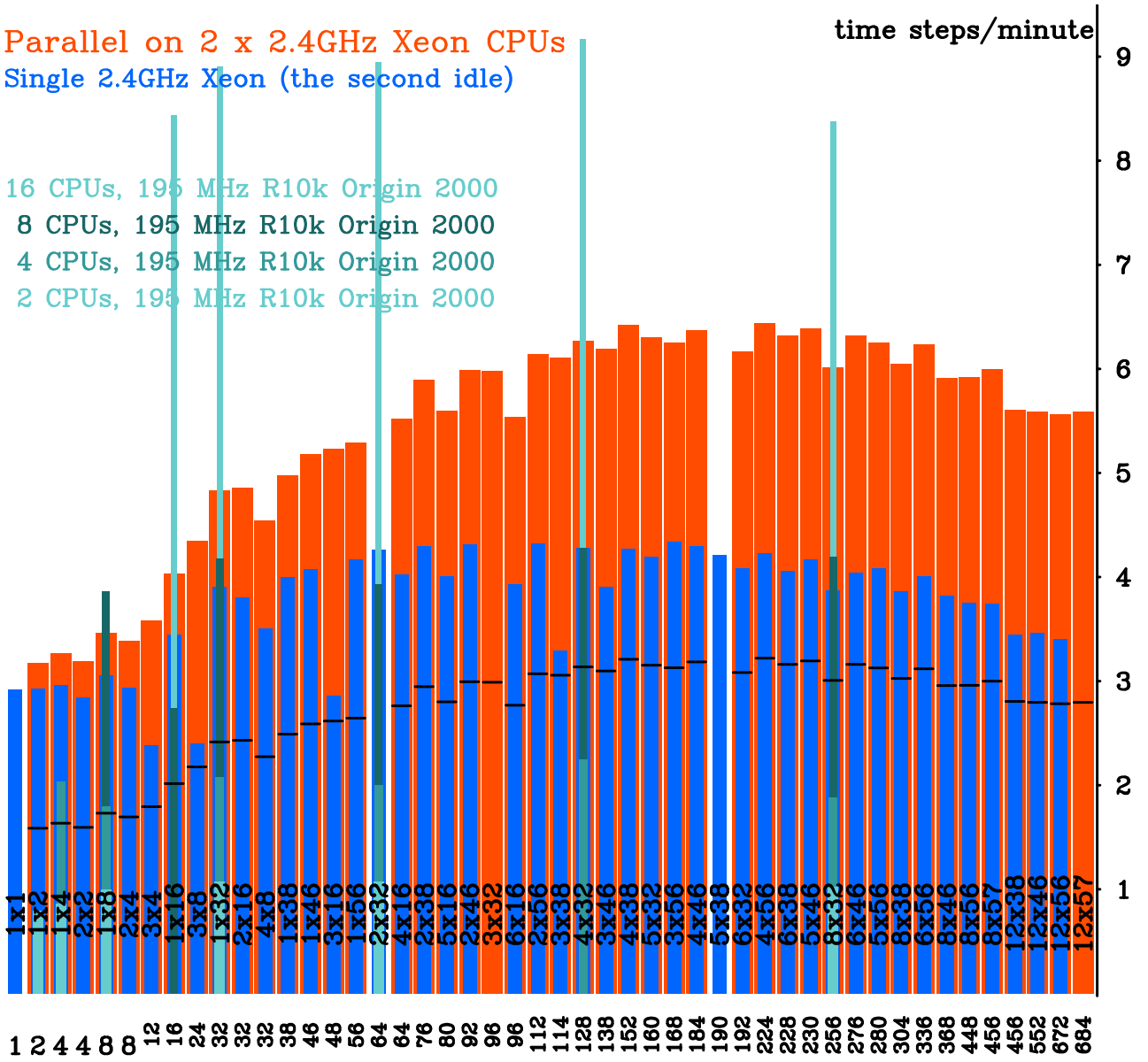
Computational performance of ROMS code as a function of subdomain partitioning (blocking) policy on different hardware platforms for 3/4 degree Atlantic model ($128 \times 128 \times 20$ grid, 100MB storage). Horizontal axis—number of subdomains (two-dimensional arrangement is written in each column in $NX \times NY$ format); vertical axis—computational performance — time steps per minute of wall clock time. In all cases parallelization is done via OpenMP and no adjustments to the code are made other than choosing different number of subdomains. Strong dependency of computation performance from number of subdomains for Intel platform is explained by cache effects due to combination of small cache, fast processors and limitation by memory bandwidth, which is by far the dominant factor for optimization strategy in this case. For all other platforms the effect is less

significant, and, in fact the most significant influence on performance can be traced to the side effects due to shortening of innermost loops when decreasing subdomain size. Nevertheless, for a properly optimized code (number of subdomains is chosen to make subdomains sufficiently small to fit into cache), even the previous generation of Intel platforms tends to outperform the other computers presented here in terms of processing power per CPU, despite the fact that its cost is only a small fraction of the cost of others.



Same as before, but showing cases with up to 16 Origin 2000 CPUs involved: the code scales perfectly to this number of processors and it is clear that it is vector length which matters most for Origin 2000

performance: blocking of the first FORTRAN dimension causes performance degradation, despite the potentially beneficial effect onto cache utilization. Sun Enterprise generally shows much lesser dependency on vector length which may be mostly explained by efficiency of Sun compilers (note 195 MHz R10k of Origin 2000 delivers the same performance as 400 MHz Sun Enterprise)



Performance 1/2 degree, $384 \times 224 \times 32$ grid Pacific model on dual 2.4 GHz Xeon machine: Overall, with proper choice of partitions the dual Xeon machine runs as fast as to 12 195MHz R10k CPUs of Origin 2000. It takes 10 hours of computing (wall clock) to get one model year of simulation, which makes it viable choice (time step 7200 sec; mode splitting ratio ndtfast=78; FB barotropic mode).

Conclusions:

Optimization strategies on Pentium PC are significantly different from that for traditional workstations and supercomputers with major accent placed on utilization of cache, and taking into account limited memory bandwidth. Length of innermost loops (i.e., vector loops) become less important for PIII, but is important again for P4 Xeon (4$\times$ longer cache line; pipelined regime)

- Linux-family operating systems and associated compilers appear to be robust and mature at this time.

- subdomain partitioning of ROMS with multiple subdomains (tiles) per processors can be used to solve cache management problem: as paradoxically as it may sound, one needs parallel code to run it efficiently on a laptop.

- For a properly optimized code 2 $\times$ 933 MHz PIIIs deliver similar (slightly better) performance as three 195 MHz R10k of Origin 2000. Dual 2.4GHz Xeon rans as fast as 12 $\times$ R10k's.

- A typical MPI code with one subdomain — one processor strategy is out of cache and would not perform/scale well on a PC

- Since 1997, when Origin 2000 was introduced, clock speed workstation-class machines increased by a factor of 3.5 or so (200 MHz R10k → 700 MHz R14k for SGI; and 330 MHz UltraSparkIIi then → 700 MHz today for Sun), while performance of PC hardware has been increased more that 10 times

- Although memory bandwidth of PCs has been improved by a factor of 6 in last three years — 500...666MB/sec of PC66...PC100 before 1999 → 1.06 GB/sec of PC133 (mainstream PIII generation PCs of year 2000) → 2.1 GB/sec of DDR PC2100 → 3.2 GB/sec RDRAM of first P4 → 6.4 GB/sec of dual-channel DDR (Intel i875/i865 sets) of today — still, modern PC are even more *miss-balanced* than their predecessors.

- Memory interleaving (common design to boost bandwidth for supercomputers) is seldom used in commodity PCs (RAMBUS of 2001 (i850/i860 sets) and dual-channel DDR of year 2003 (i875/i865) are the only 2-way interleaved system on the market today).

- Cost-performance consideration favors single-processor PCs for Linux clusters
- SMP PC designs double processor power, but memory bandwidth remains the same, which limits scalability...

... still they are capable outrun workstations and low-end supercomputers which are at least 10 times more expensive and **equalize chances of poor and rich men.**