Petascale Computing and Similarity Scaling in Turbulence

P. K. Yeung
Schools of AE, CSE, ME
Georgia Tech
pk.yeung@ae.gatech.edu

NIA CFD Futures Conference
Hampton, VA; August 2012

Supported by: NSF and NSF/DOE Supercomputer Centers, USA
The “supercomputer arms race”:

- Earth Simulator (Japan) was No. 1 in 2002 at 40 Teraflops. In 2011: the same speed did not make it into top 500.

Massive parallelism has been dominant trend

- but, because of communication and memory cache issues, most actual user codes at only a few percent of theoretical peak
- multi-core processors for on-node shared memory

Path to Exascale may require new modes of programming

Tremendous demand for resources: both CPU hours and storage

Advanced Cyberinfrastructure having a transformative impact on research in turbulence and other fields of science and engineering
Direct Numerical Simulations (DNS)

- For science discovery: instantaneous flow fields (at all scales) via equations expressing fundamental conservation laws

- Navier-Stokes equations with constant density ($\nabla \cdot \mathbf{u} = 0$):

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla \left( \frac{p}{\rho} \right) + \nu \nabla^2 \mathbf{u} + \mathbf{f}$$

- Fourier pseudo-spectral methods (for accuracy and efficiency)
  - in our work: homogeneous turbulence (no boundaries)
  - local isotropy: results relevant to high-$Re$ turbulent flows

- Wide range of scales $\implies$ computationally intensive

- Tremendous detail, surpassing most laboratory experiments
  - fundamental understanding, “thought experiments”
  - help advance modeling (both input and output)
'A $12288^3$ simulation of fully developed homogeneous turbulence in a periodic domain for 1 eddy turnover time at a value of $R_\lambda$ of $O(2000)$.'

"The model problem should be solved using a dealiased, pseudospectral algorithm, a fourth-order explicit Runge-Kutta time-stepping scheme, 64-bit floating point (or similar) arithmetic, and a time-step of 0.0001 eddy turnaround times."

"Full resolution snapshots of the three-dimensional vorticity, velocity and pressure fields should be saved to disk every 0.02 eddy turnaround times. The target wall-clock time is 40 hours."

(PRAC grant from NSF, working with BW Project Team)
2D Domain Decomposition

Partition a cube along two directions, into “pencils” of data

3D FFT from physical space to wavenumber space:
(Starting with pencils in $x$)

- Transform in $x$
- Transpose to pencils in $z$
- Transform in $z$
- Transpose to pencils in $y$
- Transform in $y$

Transposes by message-passing, collective communication

Up to $N^2$ cores for $N^3$ grid

MPI: 2-D processor grid, $M_1$ (rows) $\times$ $M_2$ (cols)
Factors Affecting Performance

Much more than the number of operations...

- Domain decomposition: the “processor grid geometry”
- Load balancing: are all CPU cores equally busy?
- Software libraries, compiler optimizations
- Computation: cache size and memory bandwidth, per core
- Communication: bandwidth and latency, per MPI task
- Memory copies due to non-contiguous messages
- I/O: filesystem speed and capacity; control of traffic jams
- Environmental variables, network topology

Practice: job turnaround, scheduler policies, and CPU-hour economics
Current Petascale Implementations

- **Pure MPI**: performance dominated by collective communication
  - usually 85-90% strong scaling every doubling of core count

- **Hybrid MPI + OpenMP** (multithreaded)
  - shared memory on node, distributed across nodes
  - less communication overhead, *may* scale better than pure MPI
    at large problem size and large core count
  - memory affinity issues (system-dependent)

- **Co-Array Fortran** (Partitioned Global Address Space language)
  - remote-memory addressing in place of MPI communication
  - key routines by Cray expert (R.A. Fiedler) on Blue Waters project, significantly faster on Cray XK6 (using 131072 cores)
DNS Code: Parallel Performance

Largest tests on 2+ Petaflop Cray XK6 (Jaguarpf at ORNL)

4096$^3$ (circles) and 8192$^3$ (triangles), 4th-order RK

CPU/step, MPI-OpenMP

pure MPI, best processor grid, stride-1 arithmetic

dealiasing: can skip some (high $k$) modes in Fourier space

better scaling when scalars added (blue, more work/core)
Future Optimization Strategies

- **Advanced MPI**: one-sided communication
  - let sending task write directly onto memory in receiving task

- **Overlap between computation and communication**
  - not a new idea, but tricky to do, and little hardware support
  - not too effective if there is not much to overlap

- **Serialized-threads**:
  - let some OpenMP threads communicate, while others compute

- **GPUs and accelerators**:
  - speed up computation and capable of very large thread counts
  - but need to copy data between GPU and CPU

- Or, shall we change the numerical method? (Consider the degree of need for communication)
Turbulence: Uses of High-End HPC

- A wider range of scales (in space and/or time)
  - higher Reynolds number (always!)
  - mixing high Schmidt number ($Sc = \nu / D$): smaller scales
  - very low $Sc$: small time steps (fast molecular diffusion)

- Improved accuracy at the small scales
  - fine-scale intermittency, thin reaction zones

- Longer simulations for better sampling or temporal evolution
  - amount of data is also a challenge

- More complex physics, coupled with other phenomena
  - e.g. stratification, rotation, MHD

- More complex boundary conditions
  - channel, boundary layer, mixing layer etc (still canonical)
Dissipation: $\epsilon = 2\nu s_{ij} s_{ij}$ (strain rates squared)

Enstrophy: $\Omega = (\nu)\omega_i \omega_i$ (rotation rates squared)

Same mean values in homogeneous turbulence, but moments and PDFs can be different

Both represent small scales, but most data sources suggest enstrophy is more intermittent, contrary to expectation at high Reynolds no. (Nelkin 1999)

Strong dissipation/straining can pull flame surfaces apart, while strong rotation leads to preferential particle concentration in multiphase flows

Difficulties in resolution and sampling, — inherent nature of infrequent but extreme events
[TACC visualization staff] $2048^3$, $R_\lambda \approx 650$: intense enstrophy (red) has worm-like structure, while dissipation (blue) is more diffuse
PDFs of Dissipation and Enstrophy

From Yeung et al. J. Fluid Mech. 2012 (Vol. 700; Focus on Fluids)

- Highest $Re$, and best-resolved at moderate $Re$ (both $4096^3$)

- High $Re$: most intense events in both found to scale similarly
- Higher-order moments also become closer
Do intense $\epsilon$ and intense $\Omega$ tend to occur together?

Yes, for most-intense fluctuations, at $R_\lambda 1000$ (and 650) (contours in first quadrant, logarithmic intervals)
Three $4096^3$ simulations have been performed, aimed at:

- Lagrangian statistics at highest $Re$ feasible
- Improved resolution of smallest scales
- Higher Schmidt number for turbulent mixing

(A fourth is planned, for mixing at very low Schmidt number)

Several hundred Terabytes of data, mostly restart files that can be analyzed to answer various physical questions

- how best to keep/organize data, at national centers
- how best to share data with other researchers
  (and/or work with them to extract statistics they need)

Cyber challenges: e.g. data management are non-trivial
Concluding Remarks

Successful extreme-scale DNS will require:

- Deep engagement with top HPC experts and vendors’ staff
- Communication, memory, and data; rather than raw speed
- Insights about the science: what will be most useful to compute, that cannot be obtained otherwise?
- Competition for hours, in high demand by other disciplines

Q.: Will we be ready for Exascale in 2018?