Some Advances in Parallel Computing

Steven Orszag
Math Department, Yale

Chief Scientific Advisor, EXA – Scalable CFD
Chief Scientific Advisor, IBRIX – Scalable I/O

Part I – Scalable CFD

Parallel CFD via Lattice BGK

- Lattice BGK methods for fluids
- Exa – roughly 70% of automotive CFD market
  - Ford, BMW, Daimler-Chrysler, Renault, Volvo, Nissan, ...
- Over 1000 licenses
- Key features
  - Highly complex geometries
  - Complex 3D separated turbulent flows
  - Fast setup – hours for a complex new car design instead of multi-weeks to months
  - Fully parallel – no 'nonlocal' pressure field needed
- Bridging the gap between engineering and design
  - Satisfying both aesthetics and engineering at once

LBM vs. CONVENTIONAL CFD

Real Fluid
Free molecules in continous space

Continuum Kinetic Theory
Microscopic particles (Boltzmann Equation)

Conventional CFD Methods

Construction of fluid equations
Navier-Stokes equations (PDE)

Discrete approximation of PDE
Finite difference, finite element, etc

Numerical integration
Solve the equations on a given mesh and apply PDE boundary conditions

Lattice Based Method

Discrete formulation of kinetic theory
Lattice Boltzmann equations

No further approximation
The equations are already in discrete form

Numerical integration
Solve on lattices and apply kinetic based BC

Simple conversion to fluid variables
These are theoretically shown to obey the required fluid equations

Results
Fluid dynamic quantities at discrete points in space and time

Lattice BGK Formulation

Lattice BGK:

$$\boxed{n_i(\vec{x}+\vec{c}_i\Delta t, t+\Delta t) = n_i(\vec{x},t) + C_i(\vec{x},t)}$$

$$C_i = -\frac{1}{\tau}(n_i - n_i^{eq})$$

Fluid quantities obtained via averaging over $\vec{c}_i$ and space-time:

$$\rho = \sum_i n_i$$

$$\rho\vec{u} = \sum_i \vec{c}_i n_i$$

$$\rho T = \frac{1}{D}\sum_i (\vec{c}_i - \vec{u})^2 n_i$$

Brief Comparison of LBGK and Conventional CFD

Conventional CFD

1. Nonlinear dynamic advection
2. Non-linear and non-local – not easily parallelized
3. Difficulties with BC
4. Geometry setup slow
5. Time dependent flows costly to simulate – especially in 3D
6. Complex physics (like multi-phase flows) require complex physical models
7. Memory – 2KB/node
8. Speed – 1μs/iteration/ node (workstation)

LBGK

1. Linear advection
2. Local and fully parallel
3. BC is fully realizable for arbitrary geometry
4. Geometry setup fast
5. Time dependent flows straightforward – especially important in 3D
6. Complex physics (like multi-phase flows) involve simple physical models
7. Memory – 0.15KB/node
8. Speed – 0.1μs/time step/ node

LBGK-τ Turbulence Modeling

- Turbulence modeled *via* a modified relaxation time τ

$$\frac{\partial f}{\partial t} = -\frac{1}{\tau}(f - f^{eq})$$

$$\frac{1}{\tau} = \frac{1}{\tau_{turb}} + \frac{1}{\tau_{shear}} + \frac{1}{\tau_{buoyancy}} + \frac{1}{\tau_{swirl}} + \ldots$$

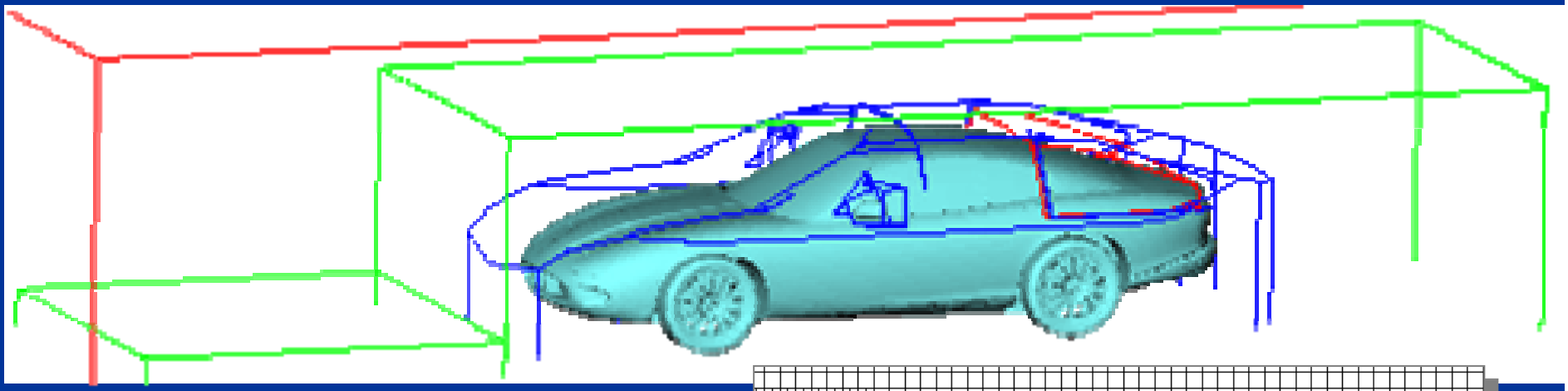
Advantages of LBGK-τ Method - I

- Realizability of the turbulence model
- LBGK-τ has guaranteed realizability
  - Requires only $\tau_{turb} > 0$
  - Stable numerical results
  - Positive eddy viscosity
- Navier-Stokes-based turbulence models can have significant difficulties with realizability
  - Divergent turbulence quantities
  - Negative eddy viscosities, …

Advantages of LBGK-τ Method - II

- The LBGK model is a higher dimensional representation of the flow in terms of space, time, and velocity as independent coordinates
- Fluid velocity/pressure are projections of the kinetic variables onto a lower-dimensional space – <u>no</u> explicit pressure needed
- Simple approximations in BGK space (like τ models) may be extraordinarily complex in fluid (velocity/pressure) variables

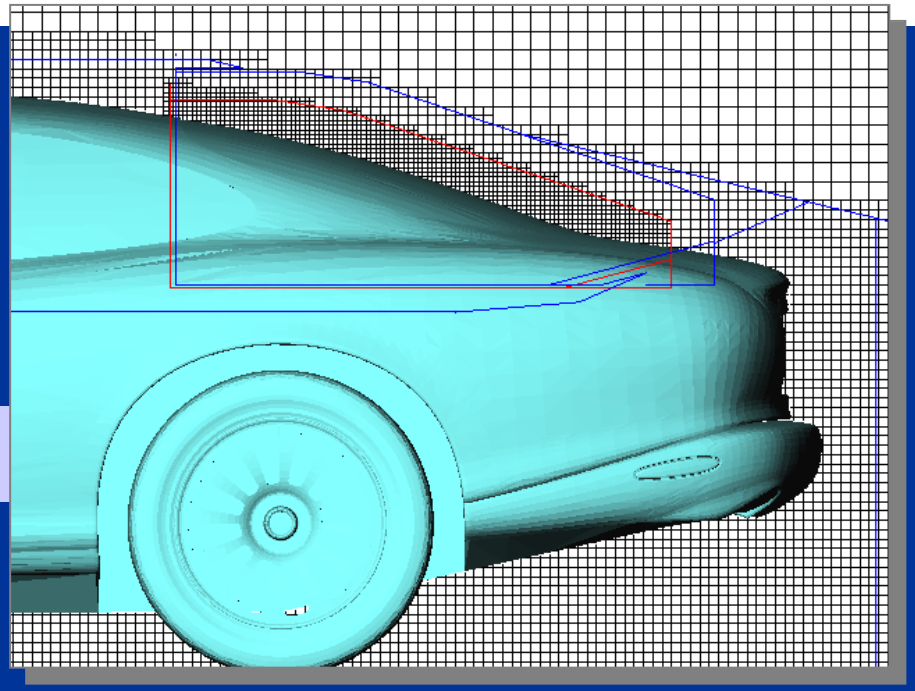
Jaguar XK8 Case Setup



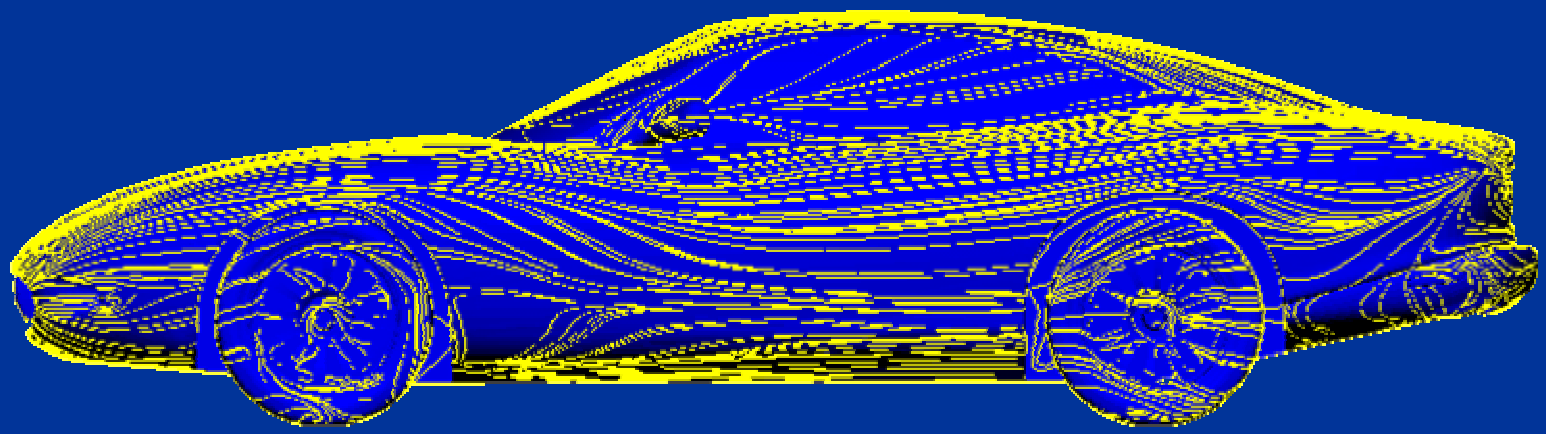
VR_5 (H44)

VR_6 (H86)

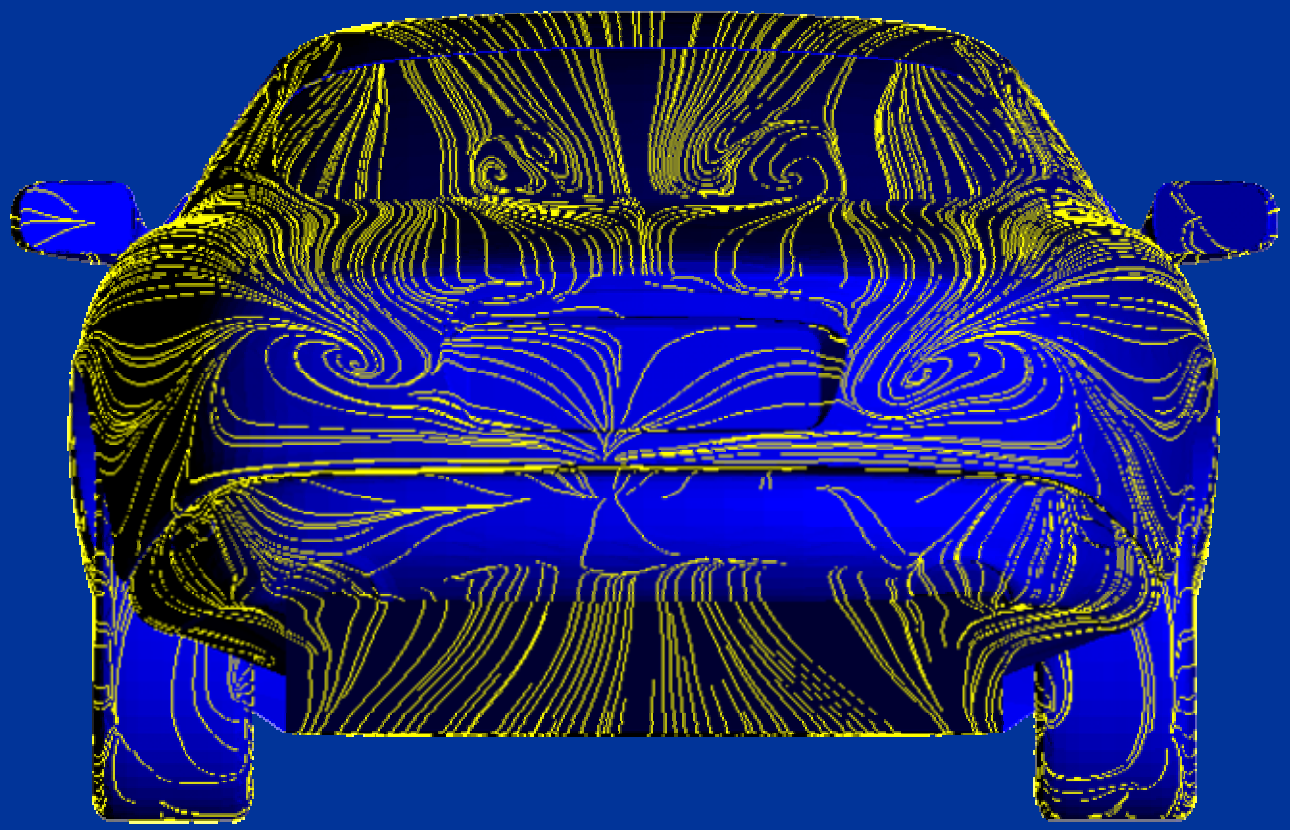
VR_7 (H192) = 5.2mm



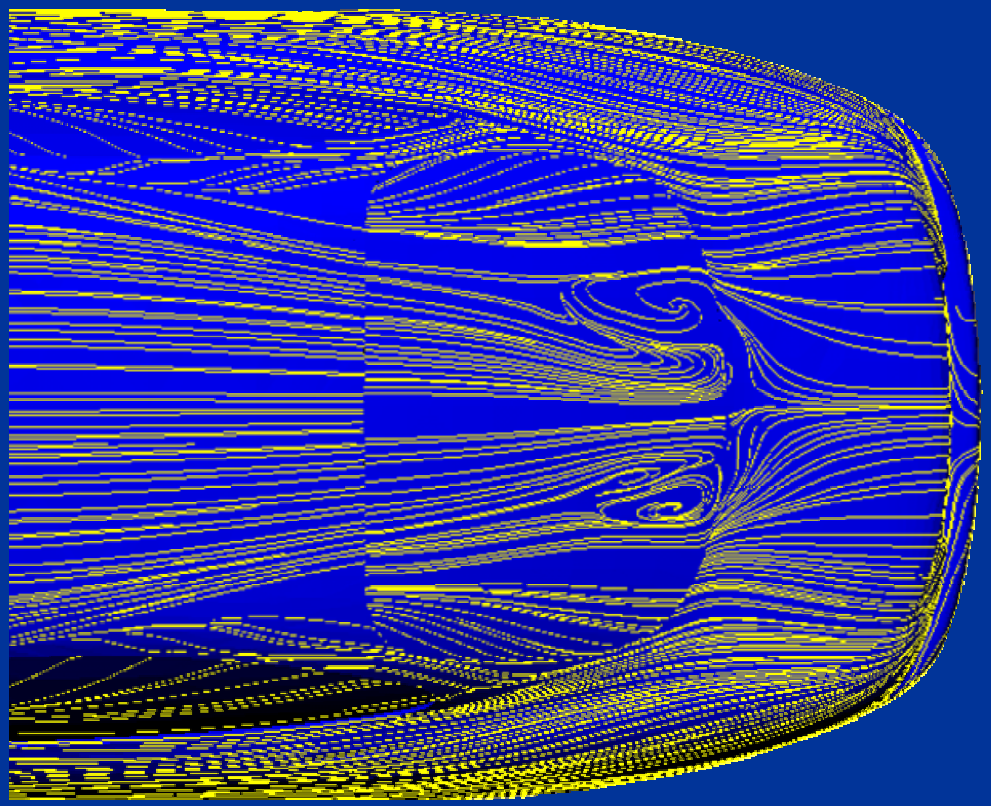
Jaguar XK8 – surface streamlines



Jaguar XK8 – surface streamlines

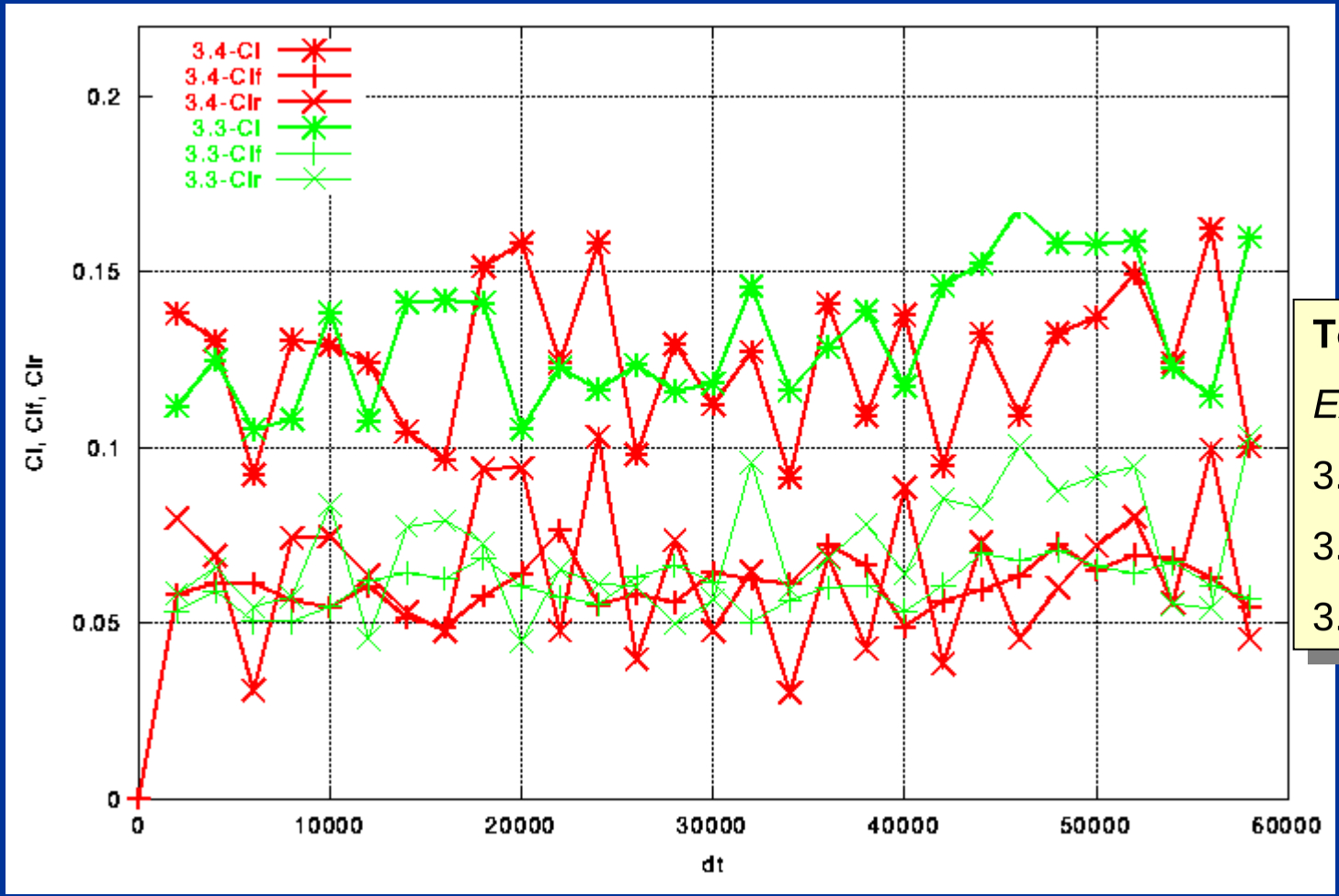


Jaguar XK8 – surface streamlines



PF 3.4

Jaguar XK8 Lift vs. Time



Total Lift:

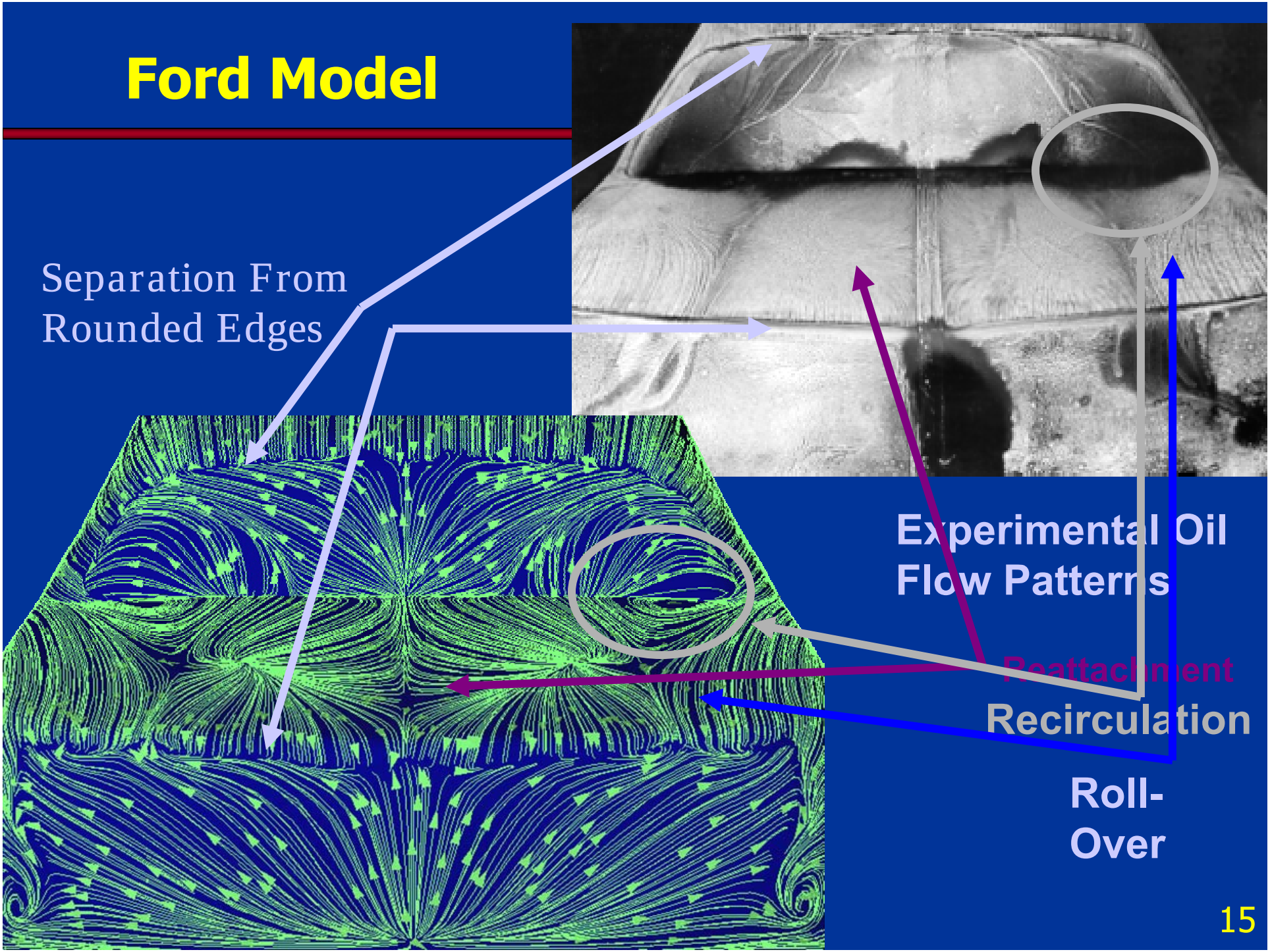
Exp: 0.128

3.2: 0.206

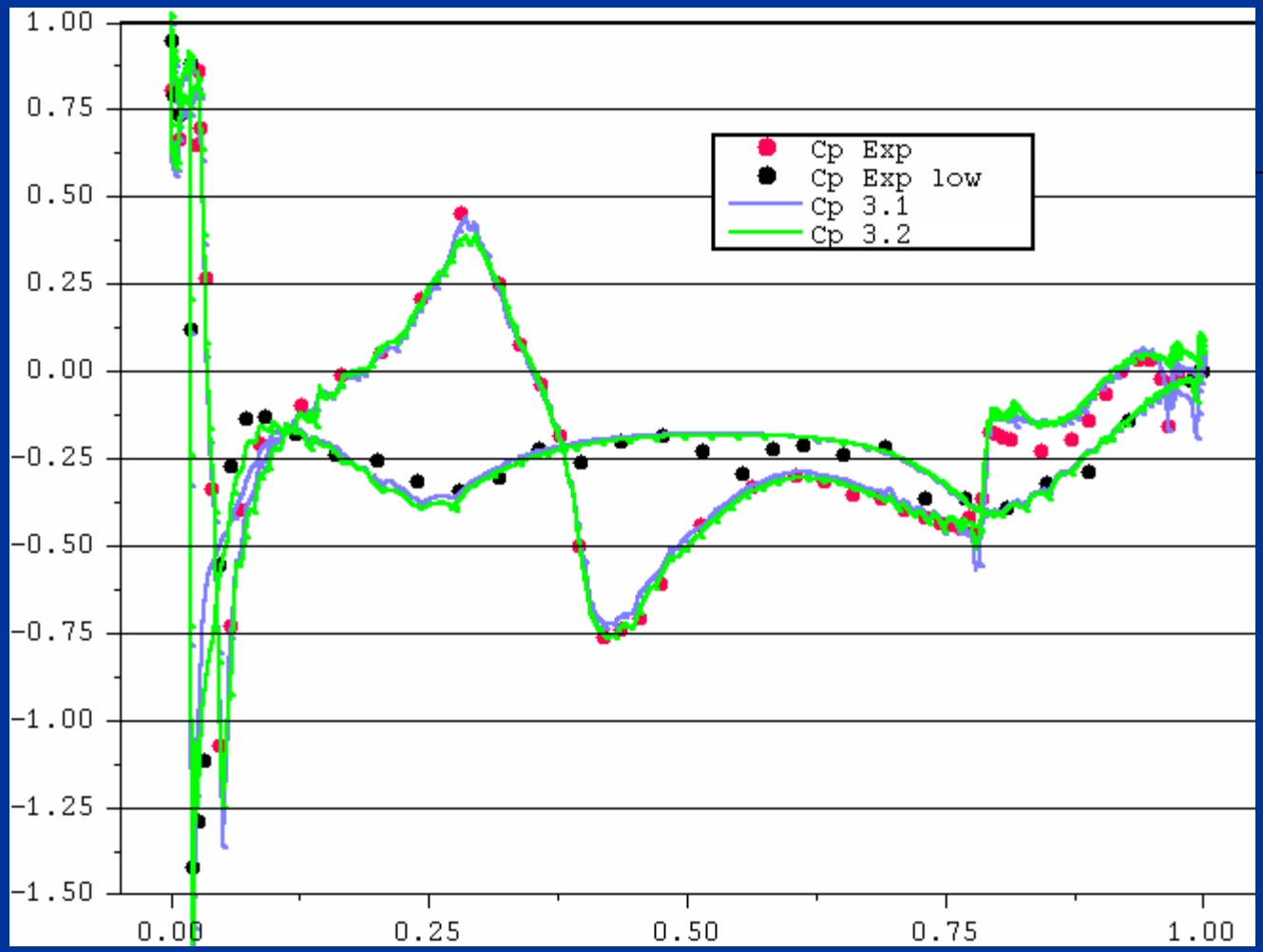
3.3: 0.146

3.4: 0.135

Ford Model


Separation From
Rounded Edges
Experimental Oil
Flow Patterns
Reattachment
Recirculation
Roll-
Over

C1 Centerline C_p (upper/lower)



Drag:

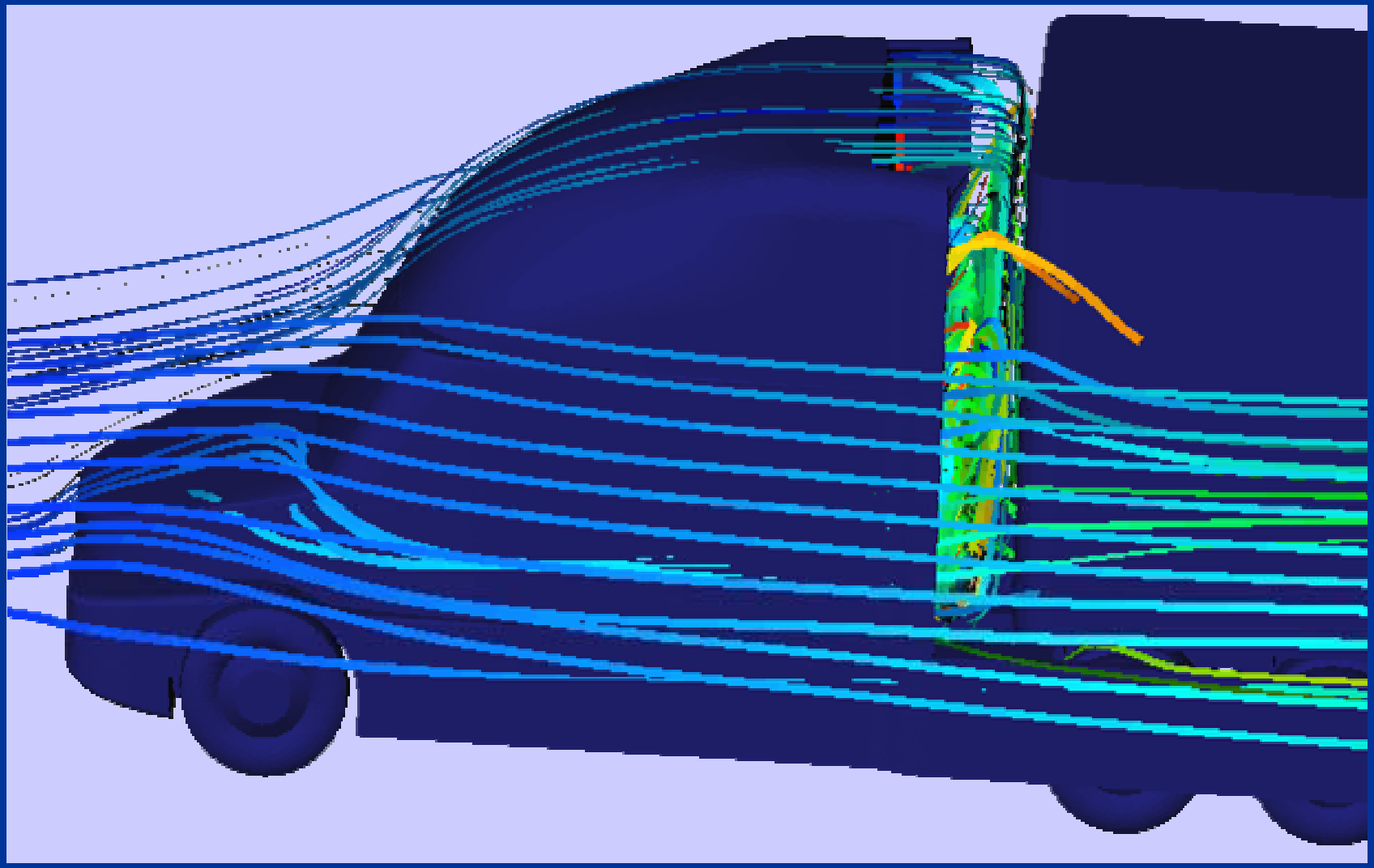
Exp: 0.161

3.1: 0.185

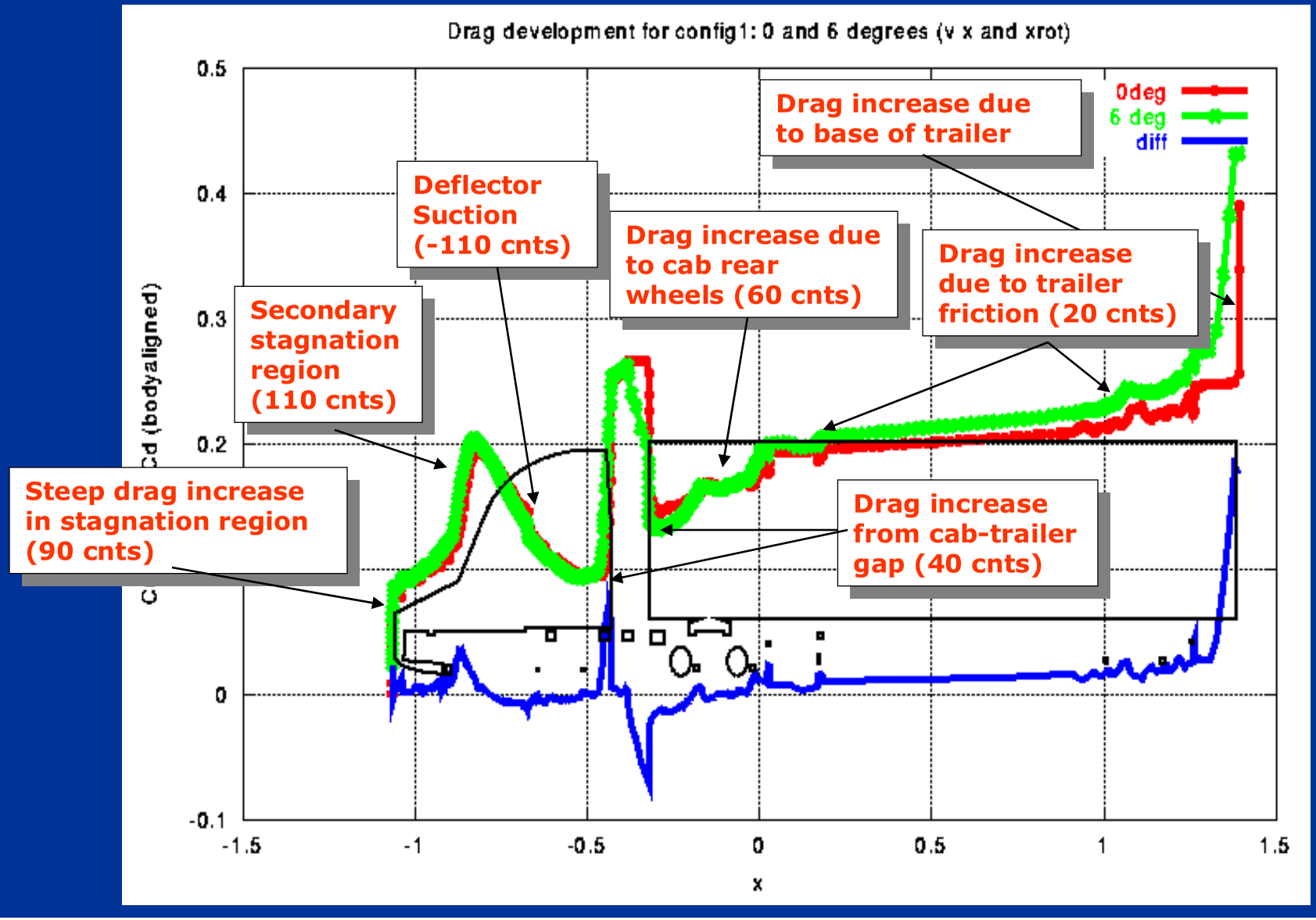
3.2: 0.160

3.3-a 0.163

3D Streamlines of Flow Past a Large Truck

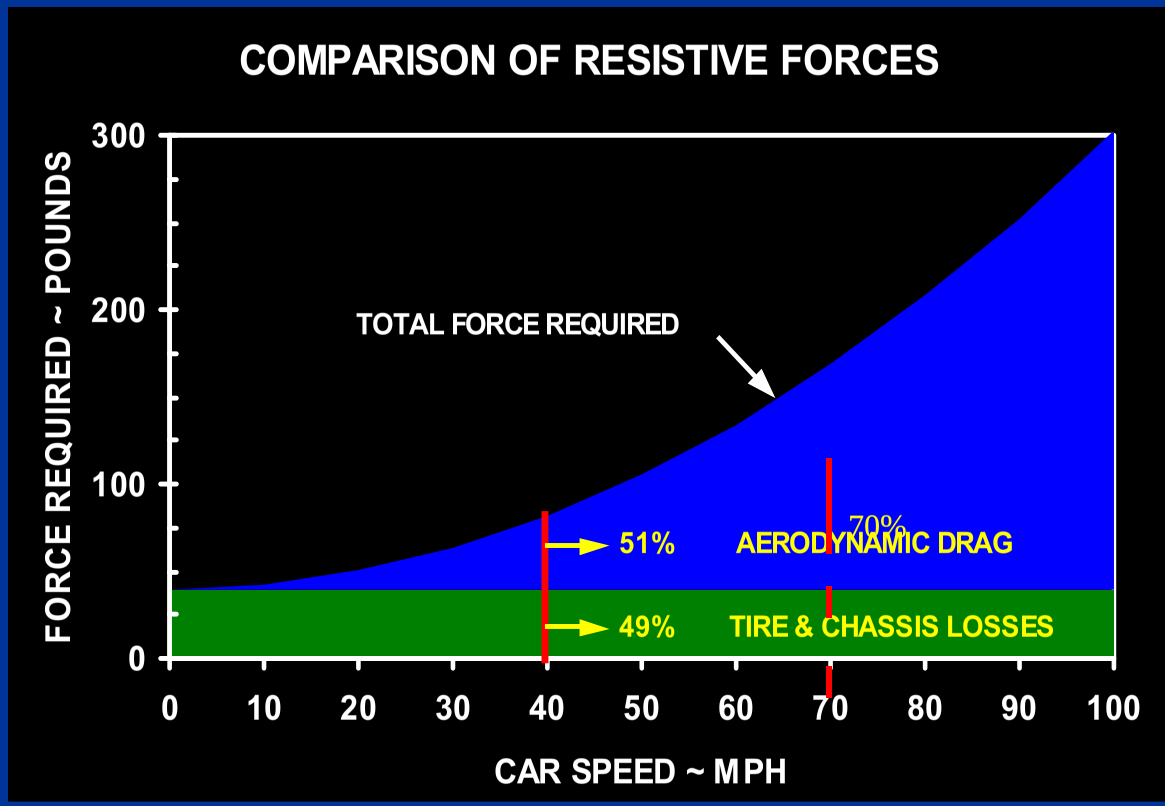


Drag Development



Fuel Economy

- CAFE Standards
  - Being raised and definitions of 'light truck' changing
  - 50% to 75% of resistive highway forces due to aero drag
- Value of 1 MPG
  - CAFE = $55/vehicle
  - x 1.2M vehicles
  - $66M for Explorer
  - Over 4 year life

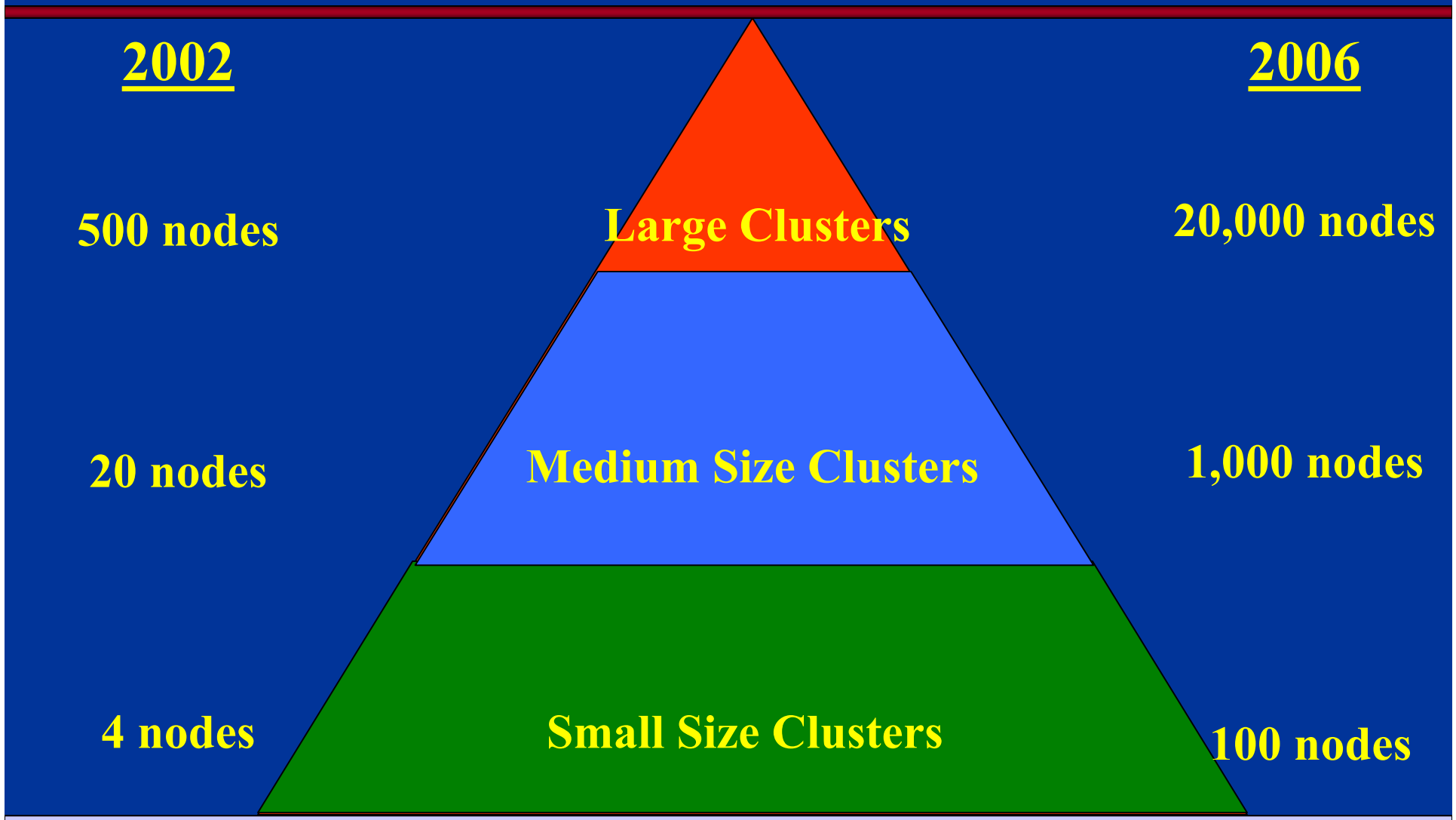


Conclusions – Part I

- **LBGK-τ two-equation model describes complex large-eddy structures**
- **Some critical issues:**
  1. Time dependence is critical for high Reynolds number 3D flow separation
  2. Higher effective Reynolds numbers of LBGK-τ is critical for heat transfer, separation, and swirl/rotation
  3. Transition/intermittency modeling in VLES critical due to higher effective Reynolds number
  4. VLES critical for proper coupling of buoyancy, rotation, heat release, ...
- **LBGK-τ combines these advantages into a useful tool for real-world CFD**

Part II – Scalable I/O

Challenge: Clusters are Getting Bigger



File systems designed for the large clusters of 2002 cannot support even next-gen medium-size clusters!

Challenges of Next-Generation Cluster Data I/O

(as well as Next-Generation Grid and Enterprise Computing)

Scalability	
✔ Capacity – single namespace	✔ Performance – high throughput
✔ Distributed files/directories	✔ Large and many small files
Adaptability	
✔ Responsive allocation policy	✔ Varying access patterns
✔ Changing workloads	✔ High and low write/read ratios
Availability	
✔ No single points of failure	✔ Partial off-line repair of FS
✔ Dynamical load balance	✔ Dynamical recovery
Manageability	
✔ Simple, centralized mgmt	✔ Easy to upgrade/use

IBRIX Single Namespace Distributed File System

- Unprecedented scalability in performance and capacity
  - Fully symmetric loosely coupled distributed metadata
- High availability
  - File system operational/repairable even if parts are offline
- Unprecedented flexibility and adaptability
  - Scalable NAS and/or cluster file system
  - Uses any combination of DAS and/or SAN-attached storage
  - Hardware/protocol agnostic
- Fully integrated product
  - File system, LVM, HA, Management (GUI/CLI)

IBRIX *fusion* – Features and Benefits

Scalability	
❖ Distributed metadata management	✔ Scales linearly to 1000s of servers with no single-MDC bottleneck
❖ Non-hierarchical placement	✔ Highest aggregate performance
❖ Control migration	✔ Load balancing
❖ Direct data access over SAN	✔ Highest performance even with heavy file read sharing
❖ Distributed files	✔ High single file throughput

Performance	
❖ Sophisticated lock management and cache coherence	✔ No performance degradation vs. local file system
❖ Fixed overhead protocol for remote metadata access	✔ Highly efficient for both small and large block I/O

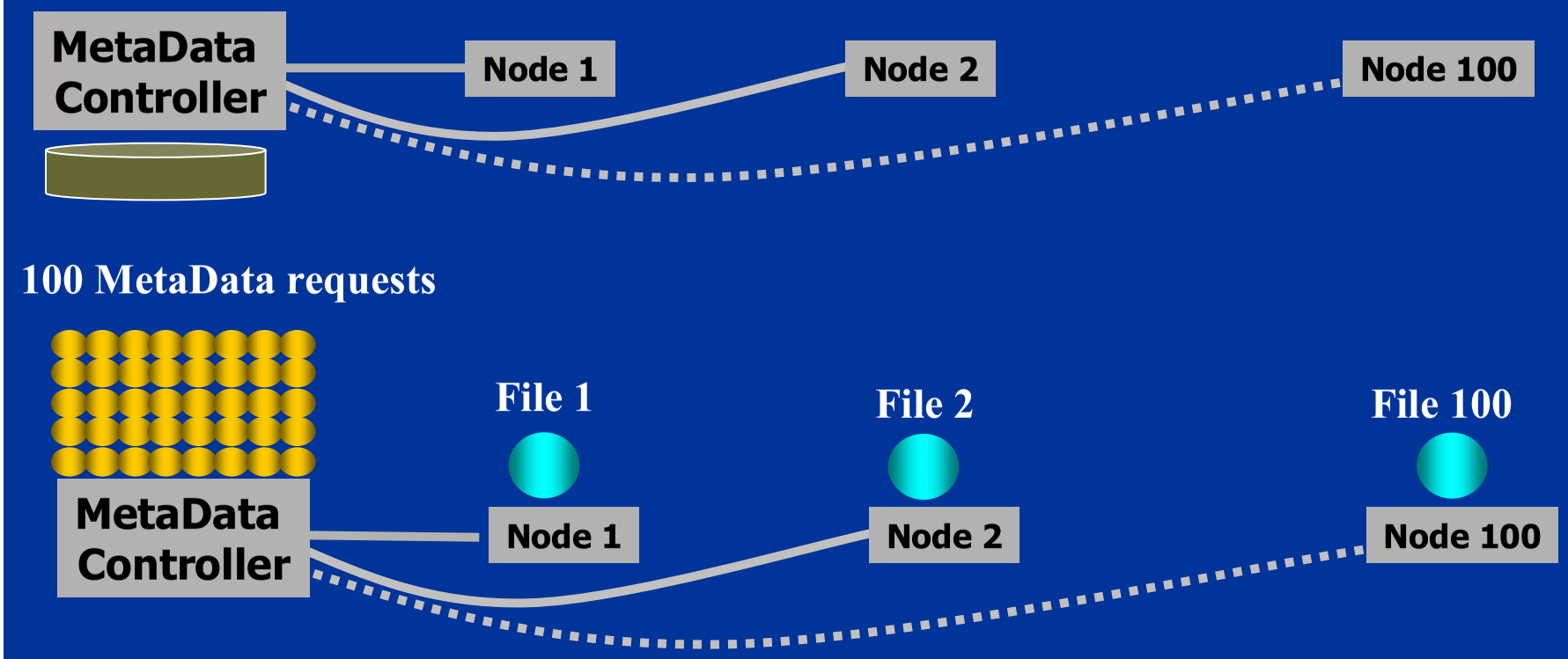
IBRIX *fusion* – Features and Benefits

Availability	
❖ Automatic failover	✔ Uninterrupted operation
❖ Partial FS operation	✔ Fault containment for min downtime
❖ Control migration	✔ Flexible maintenance

Manageability	
❖ Cluster-wide configuration via web-based GUI and CLI	✔ Simple and comprehensive centralized management
❖ Standards based: XML, SNMP	✔ Compatible with 3rd party software
❖ Performance statistics, traps, and event notification	✔ Self-contained

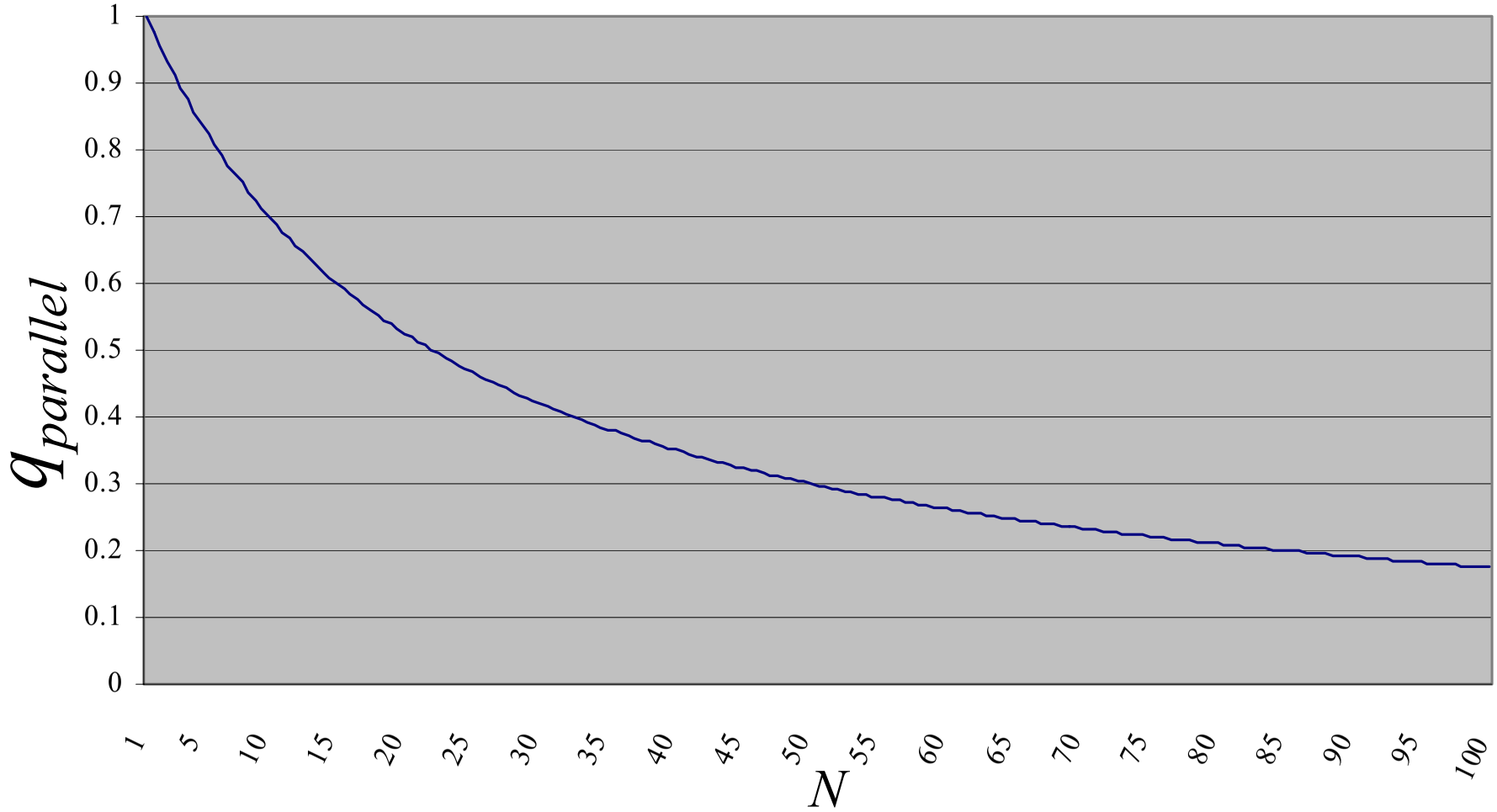
Old 'solution': Out-of-band metadata controller

Single Metadata Controller for File System



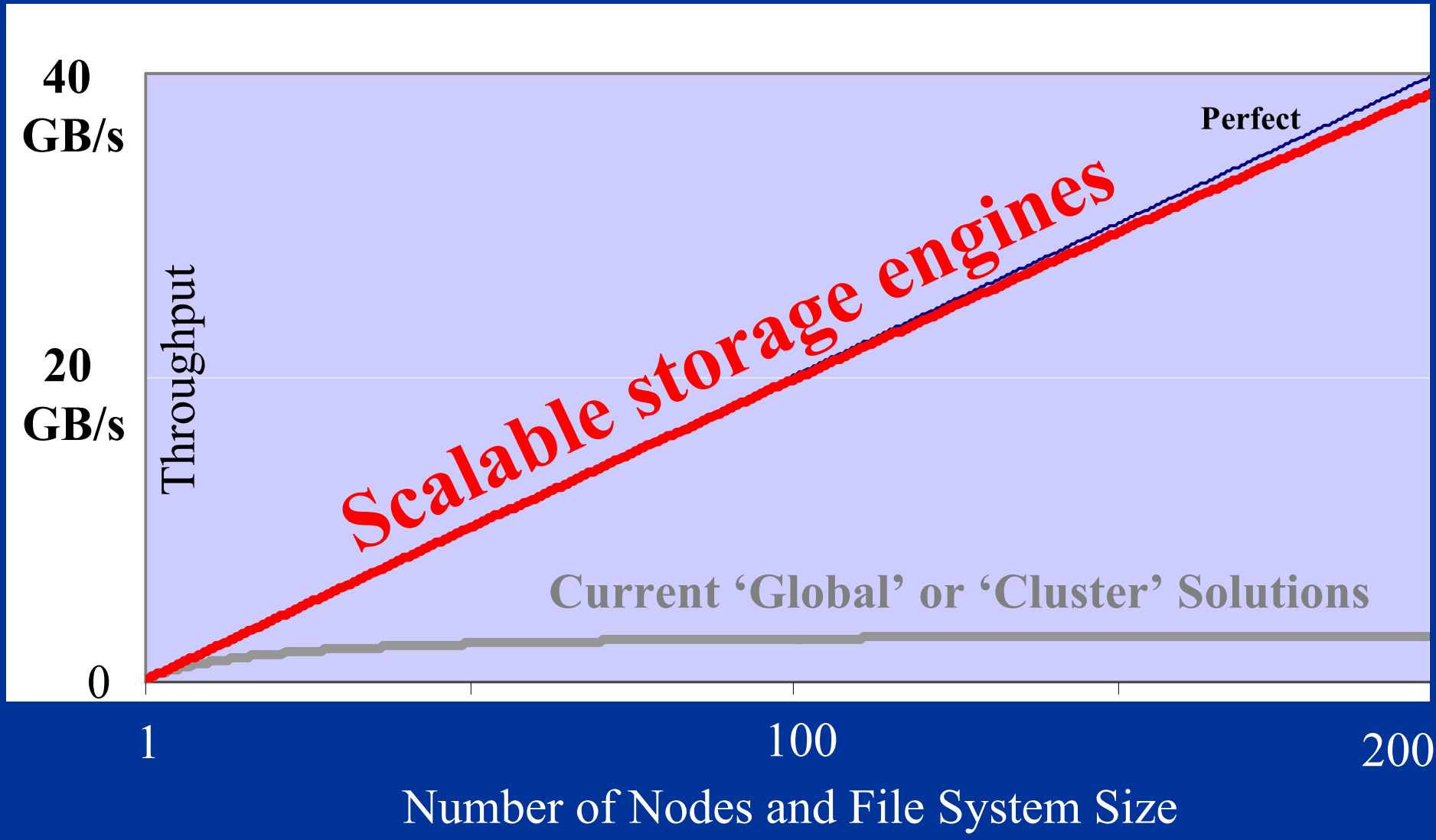
Conclusion: MetaData controller is a bottleneck

Parallel Performance of Out-of-Band Metadata Control (f =MD/D= 5%)



$$q_{parallel} = \frac{1+f}{1+Nf}$$

Throughput vs. Number of Nodes



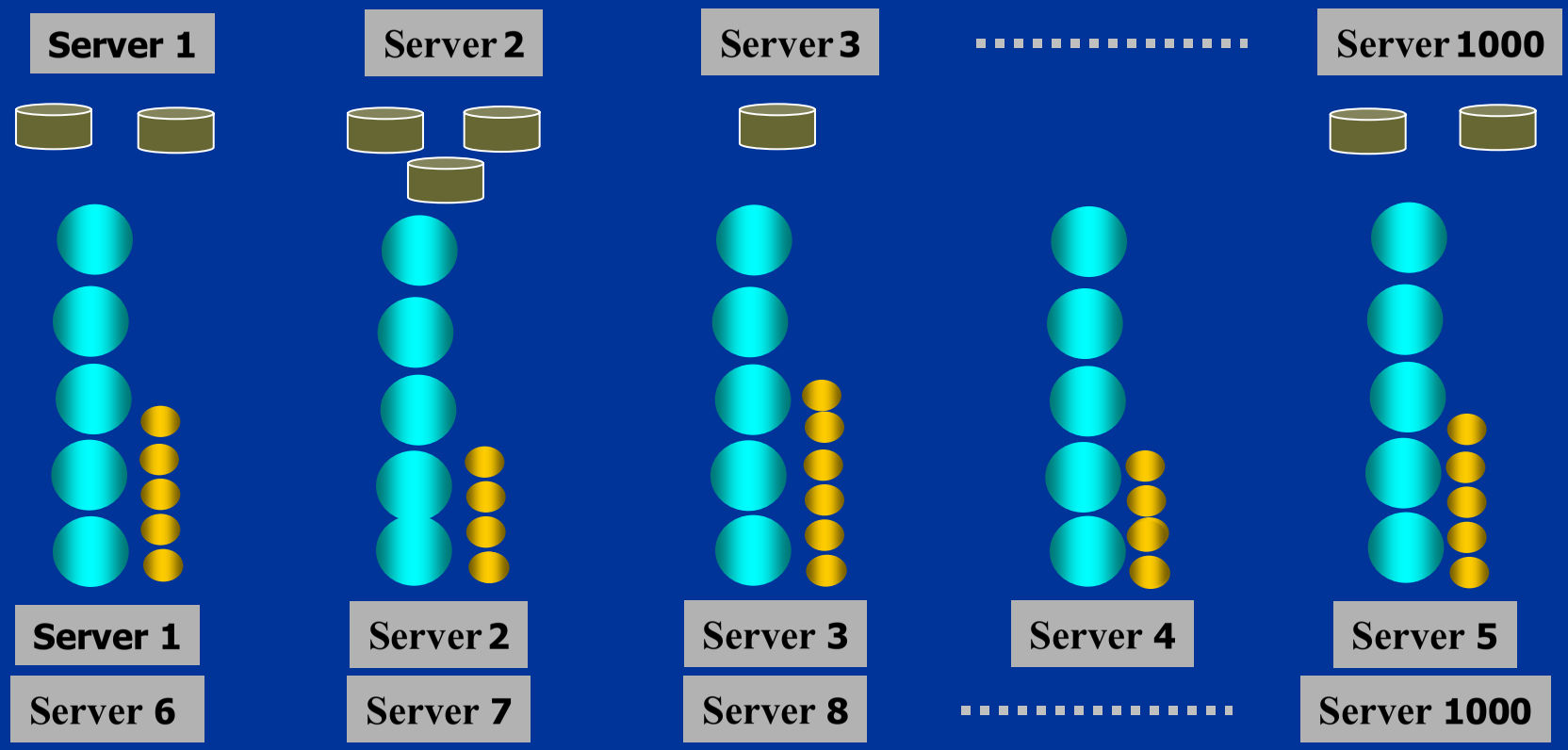
Classification of Data Access Patterns

1. Embarrassingly Parallel (EP) – clients attached to each storage engine request locally controlled data
2. Random access – clients request data in a completely random way – files that are requested have metadata that is controlled by a random storage engine
3. Group (or clustered) access – Random access, but only among subgroups or clusters
4. Hot-spot access – 'Hot' files are frequently accessed

- Enterprise storage routing architecture
  - EP perfectly parallel – 100% efficiency
  - Random – perfectly parallel with probability 1 – nearly 100% efficiency
  - Group – perfectly parallel with probability 1 – nearly 100% efficiency
  - Hot-spot – Scalable performance *via* novel cache design – high efficiency

Segmented File System

Distributed Metadata Control



Conclusion: MetaData is NOT a bottleneck
Nearly 100% scalability

Theory of Perfect Scalability
'Extreme Statistical Mechanics'

- 'Toy' mathematical model
  - Deposit *kN* balls randomly into *N* boxes
  - What is the expected value of the **maximum** number of balls in a single box?
- For $k \to \infty$, we obtain the 'steady state' behavior

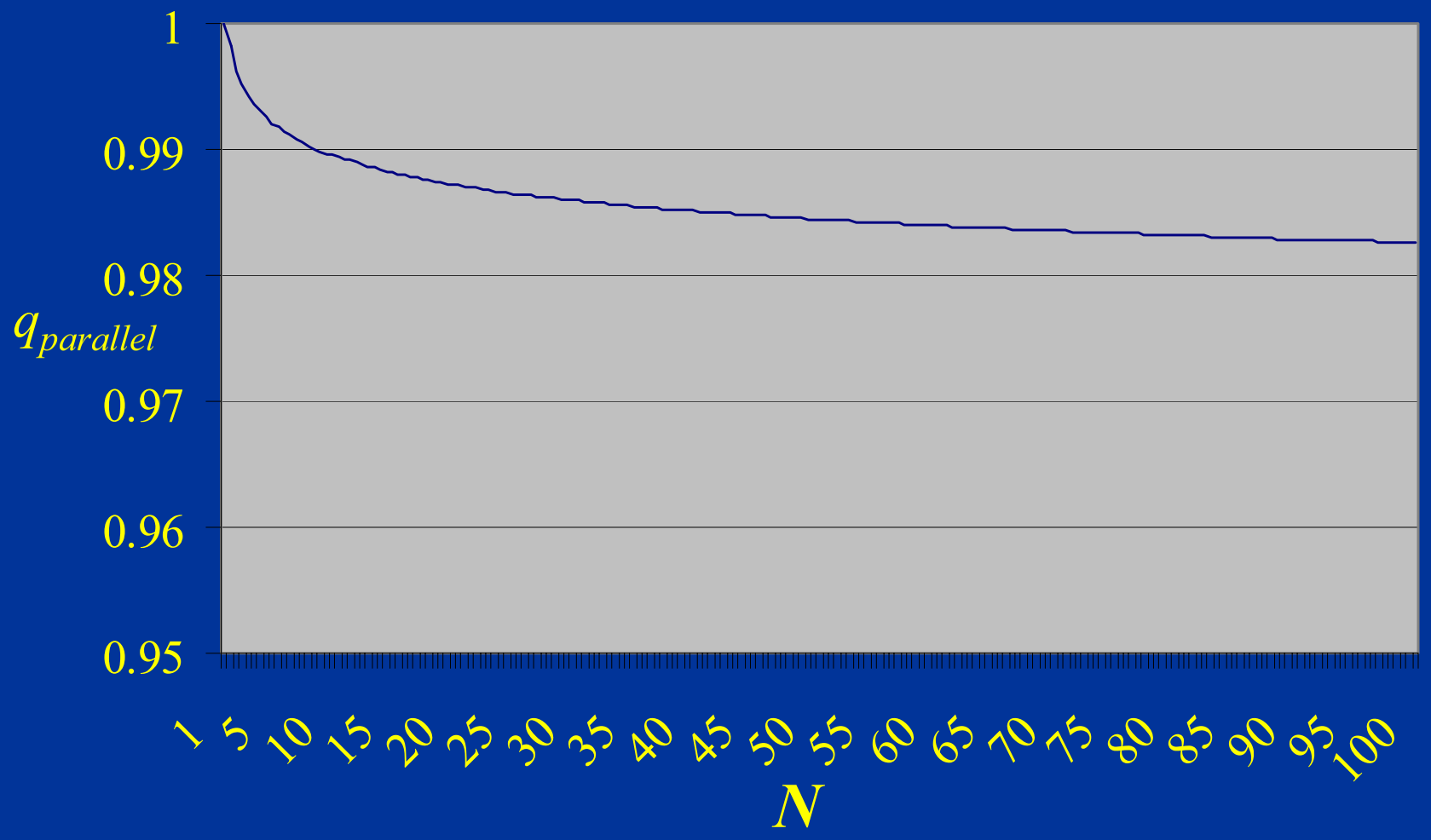
$$\frac{\langle n_{\max} \rangle}{k} \approx 1 + \sqrt{\zeta(N)\left[\frac{2}{k} + \frac{\zeta(N)}{k^2}\right]} \to 1$$

$$\zeta(N) \equiv \xi(N) - c$$

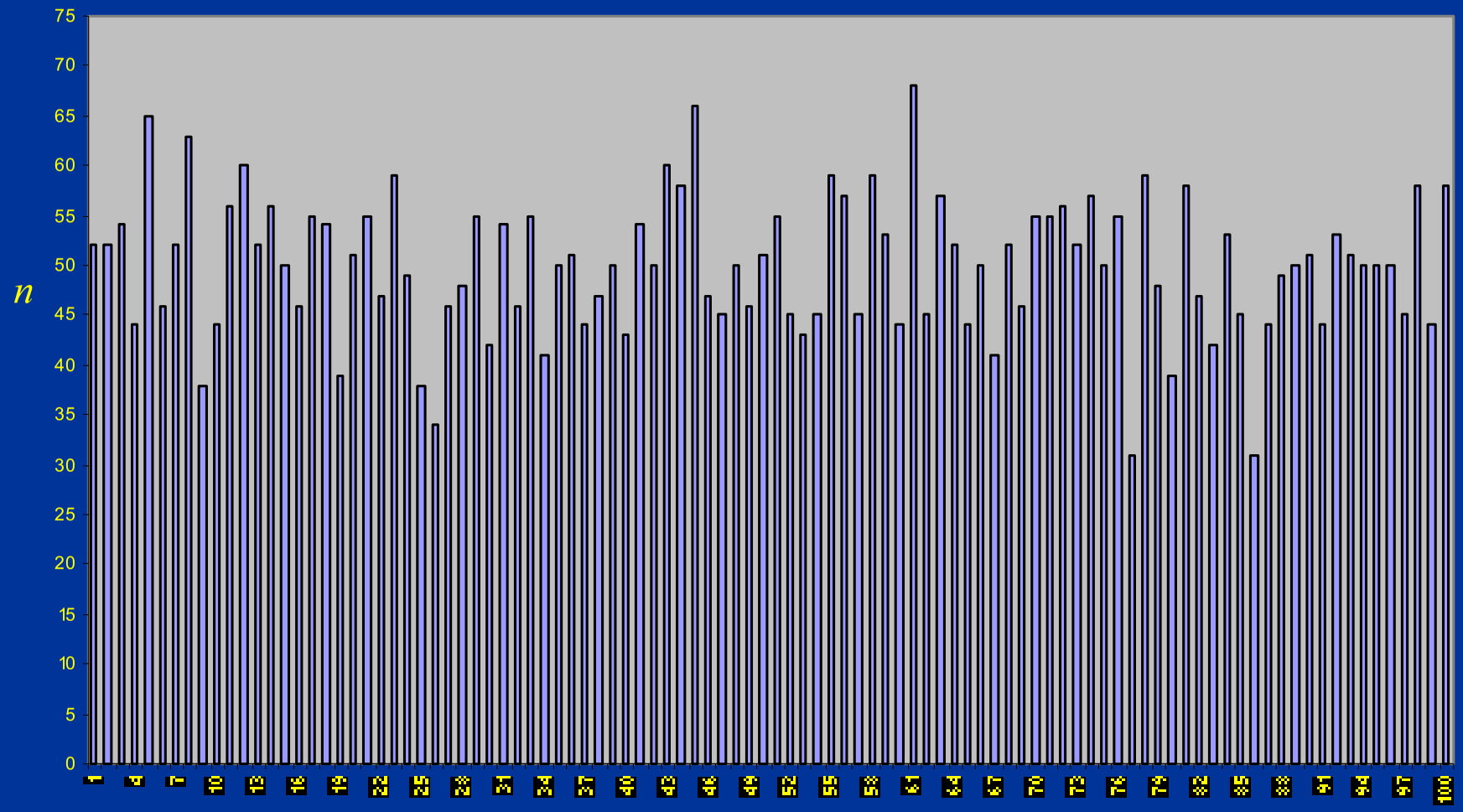
$$c \approx 1.66$$

$$\Gamma(\xi(N)) = N, \quad \xi(N) \sim \frac{\log N}{\log\log N}(1 + o(1))$$

Parallel Performance (50 requests/server)



Steady State Case: Metadata Queues
50 requests/server, 100 servers



Conclusions – Part II

- Scalable I/O is achievable today and can meet the challenges posed by effective utilization of next-generation computing systems
- Scalability can be achieved without sacrificing key enterprise features (HA, manageability, …)
- These new technologies promise significant improvements in both high-end HPC environments as well as enterprise and SMB data environments