

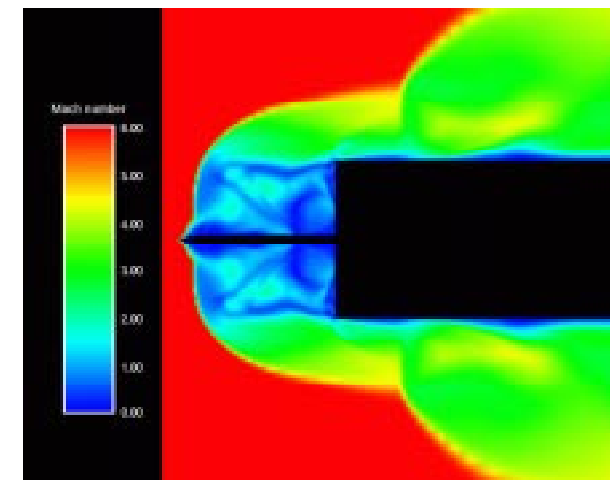
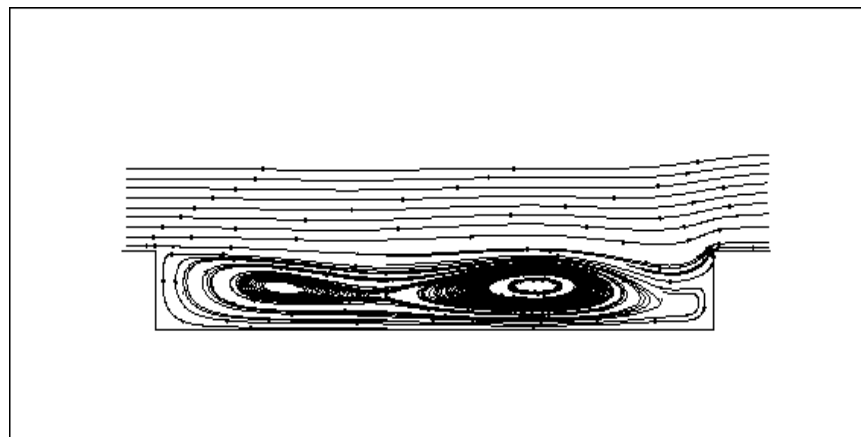
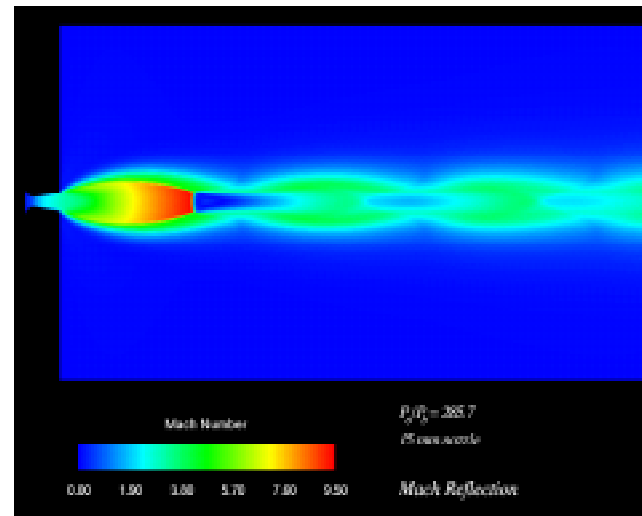
# Beowulf Clusters

K.J.Badcock, M.A.Woodgate, K.Stevenson,  
B.E.Richards.

Computational Fluid Dynamics Group,  
University of Glasgow

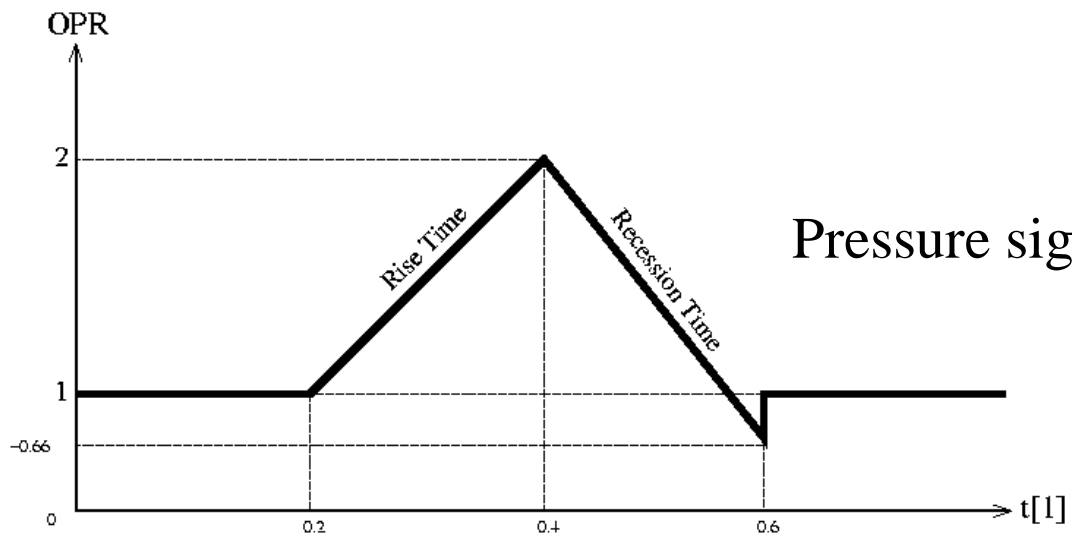
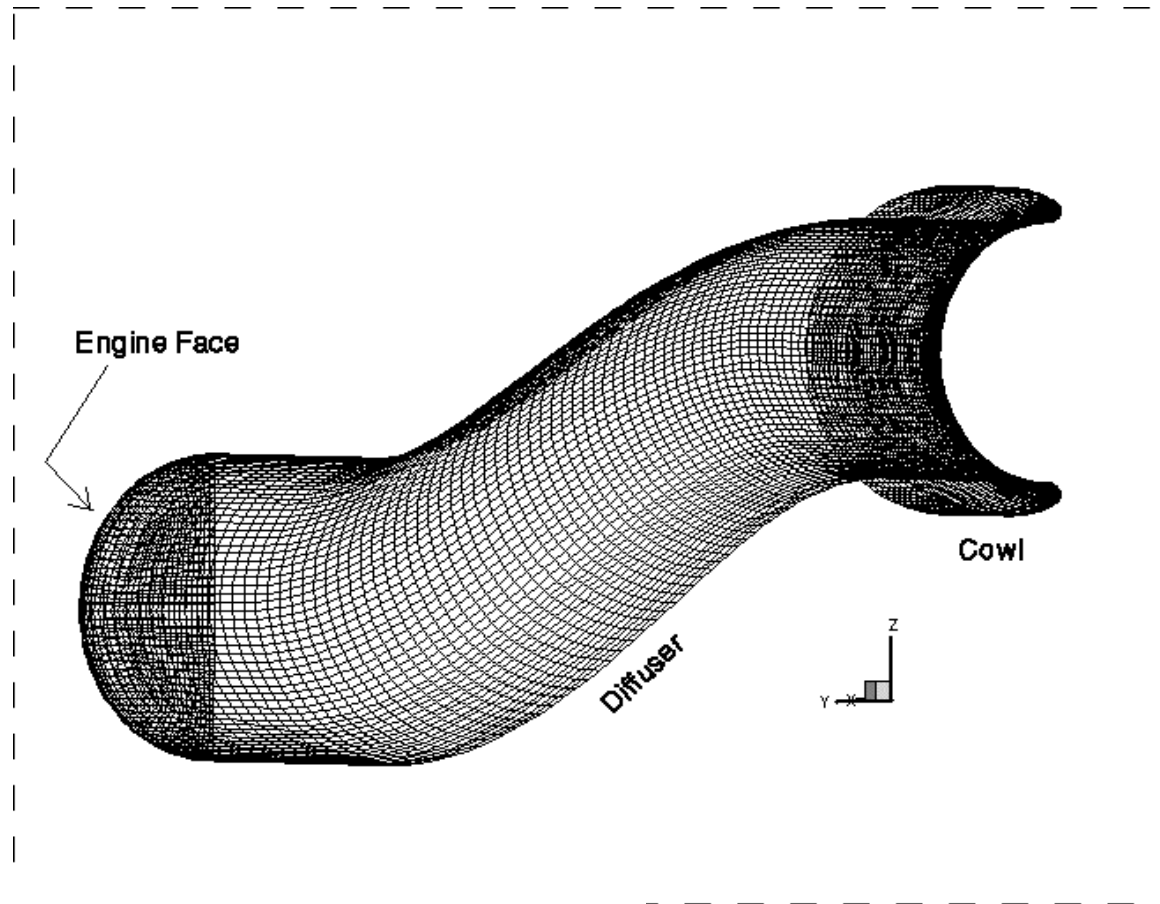
<http://www.aero.gla.ac.uk/Research/CFD>

# 16 node P200 cluster, 1997

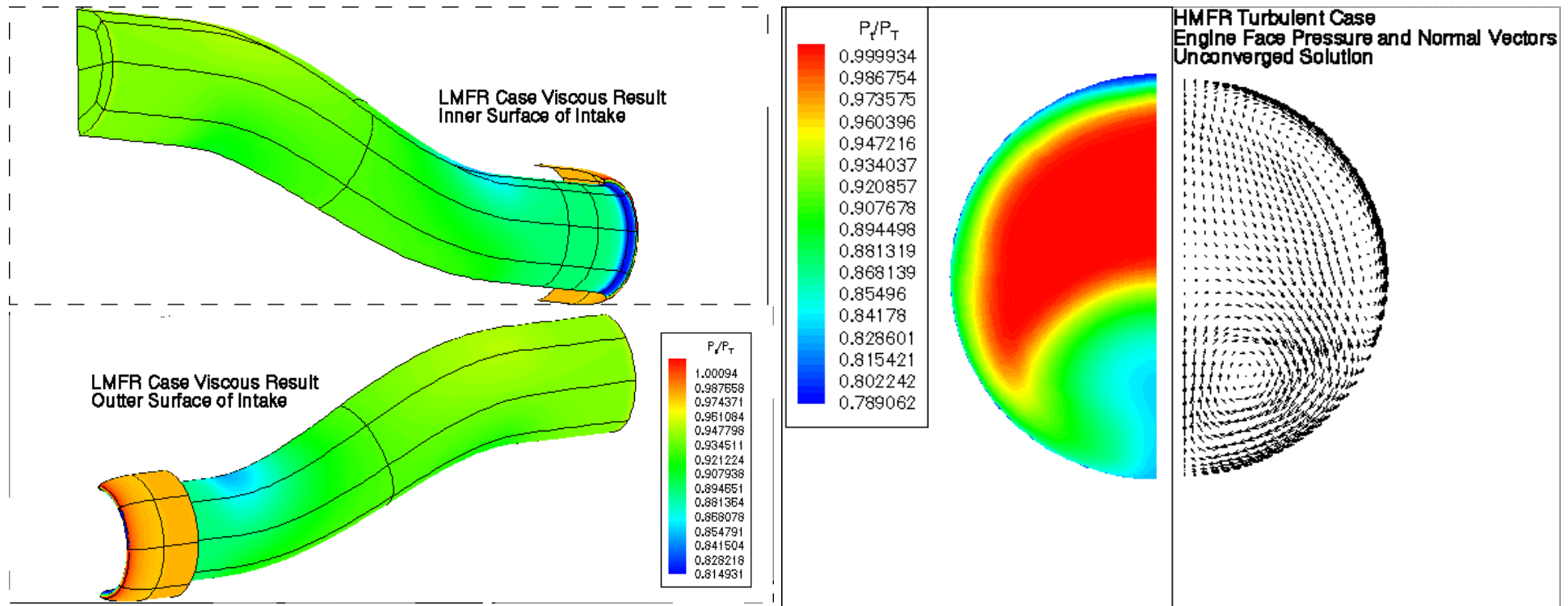


# Numerical Method

- Spatial discretisation
  - Osher or Roe for convection, CD for viscous
- Turbulence Modelling by k-w
- pseudo time stepping
  - iteration by unfactored implicit method
  - Krylov subspace method
  - blocked BILU preconditioning
- Parallel implementation using MPI

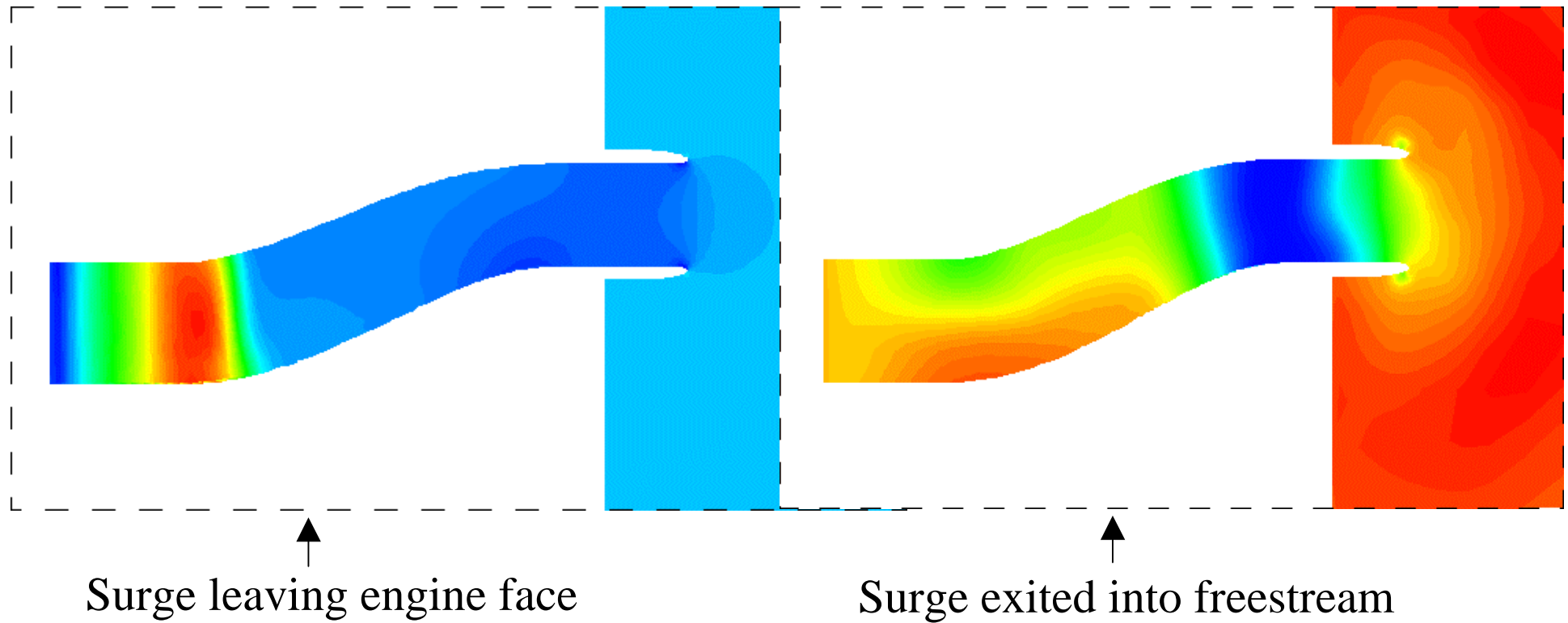


Pressure signature applied to engine face



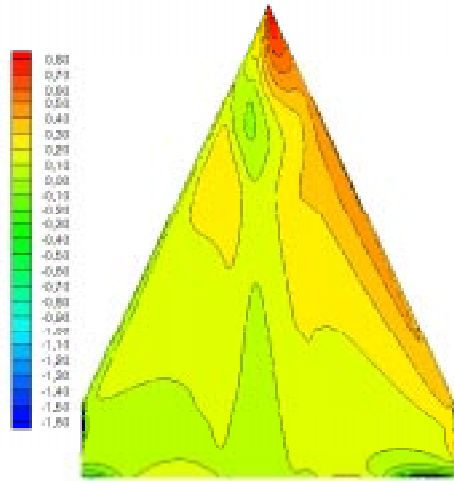
## Steady Validation

- Validation against previous computations and experimental work was satisfactory for low and high engine demand
- Substantial secondary flow can be seen to be generated at the engine face, particularly for the high engine demand case
- Medium grid (400,000+ GP's), 8 processors, 10 hours elapsed calculation time

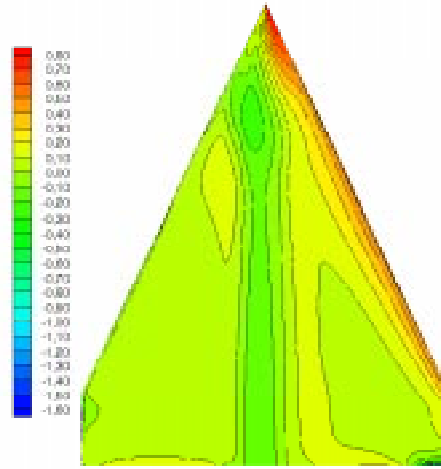


- Medium grid (400,000 GP's)
- Navier Stokes Calculation (k- $\omega$  turbulence model)
- $Dt=0.005 \Rightarrow$  1000 iterations
- 8 processors  $\Rightarrow$  11 hours elapsed calculation time

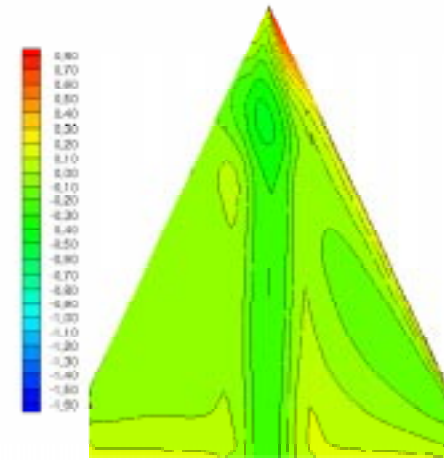
lower



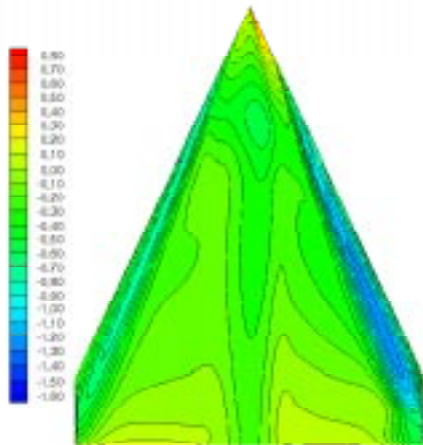
30



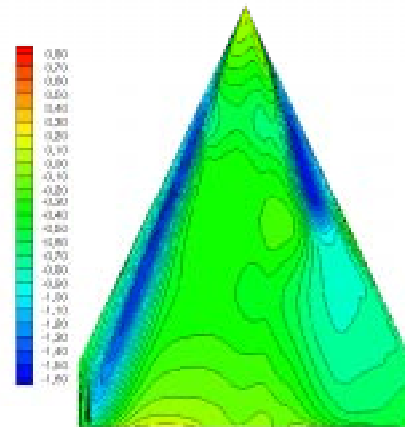
60



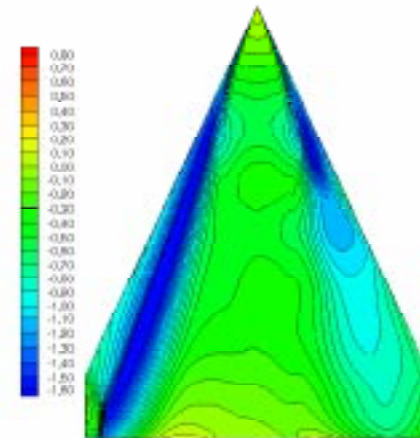
90



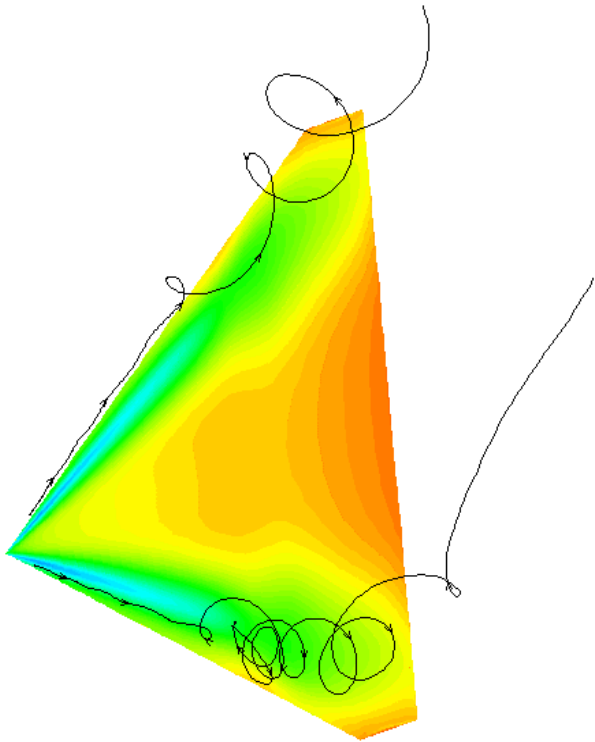
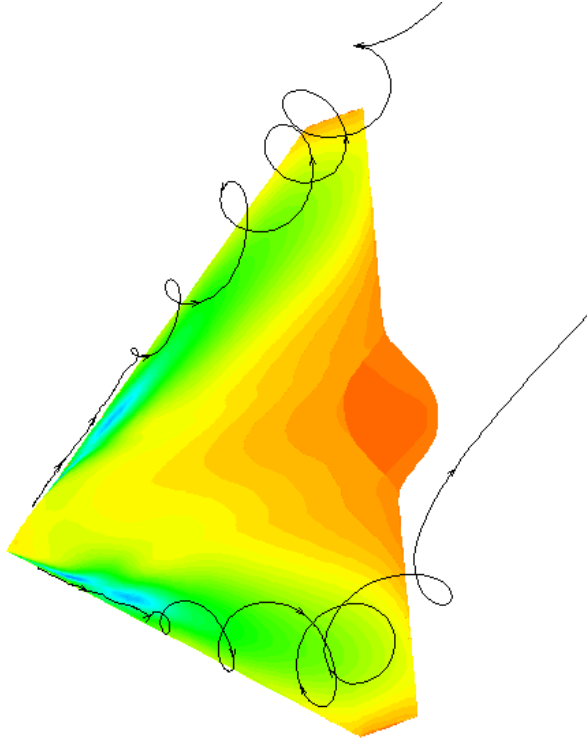
120



150



180



# Steady Calculation Performance

- Fine Grid with  $> 3,000,000$  Grid Points
  - 500 implicit steps
  - 15 Processors
    - » 4 Hours
- Medium Grid with  $> 400,000$  Grid Points
  - 300 implicit steps
  - 8 Processors
    - » 0.5 Hours

# Unsteady Calculation Performance

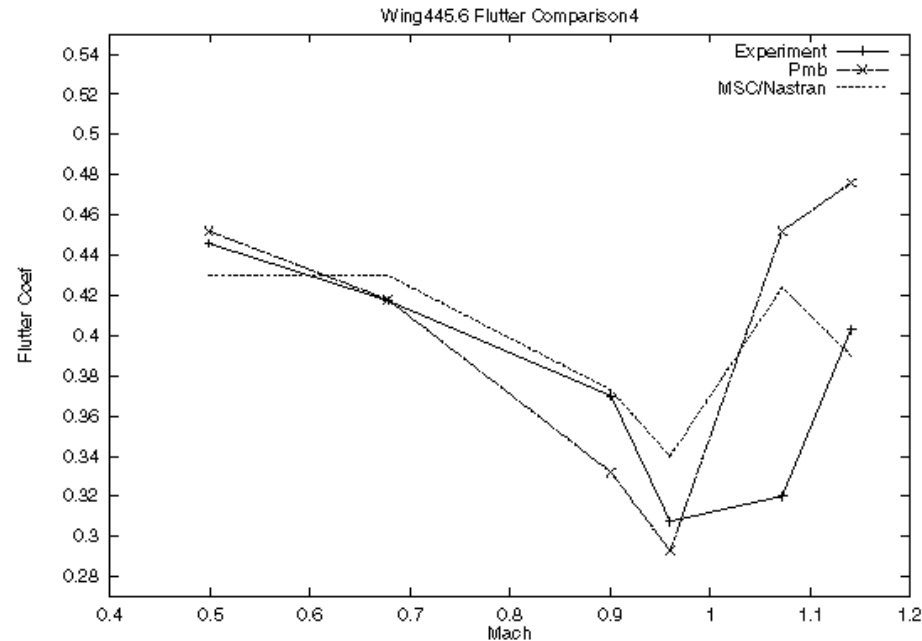
- Medium Grid with  $> 400,000$  Grid Points
  - Pitching Calc -3 cycles @ 50 time steps / cycle
  - 8 Processors
    - » 1.5 Hours
  - Yawing Calc - 6 cycles @ 200 time steps / cycle
  - 8 Processors
    - » 11.5 Hours

# Flutter Calculations

- Coupled flow code with modal structural model
- data transfer methods evaluated
  - [www.aero.gla.ac.uk/Research/CFD](http://www.aero.gla.ac.uk/Research/CFD)
- 2 applications
  - flutter boundaries through time marching calculations
  - aerostatic deflections



- MDO wing aerostatic calculation
- $M=0.88$ , incidence= $-0.25$  degrees
- 300k points on 2 processors
- 2.4 hrs to converge 5 orders



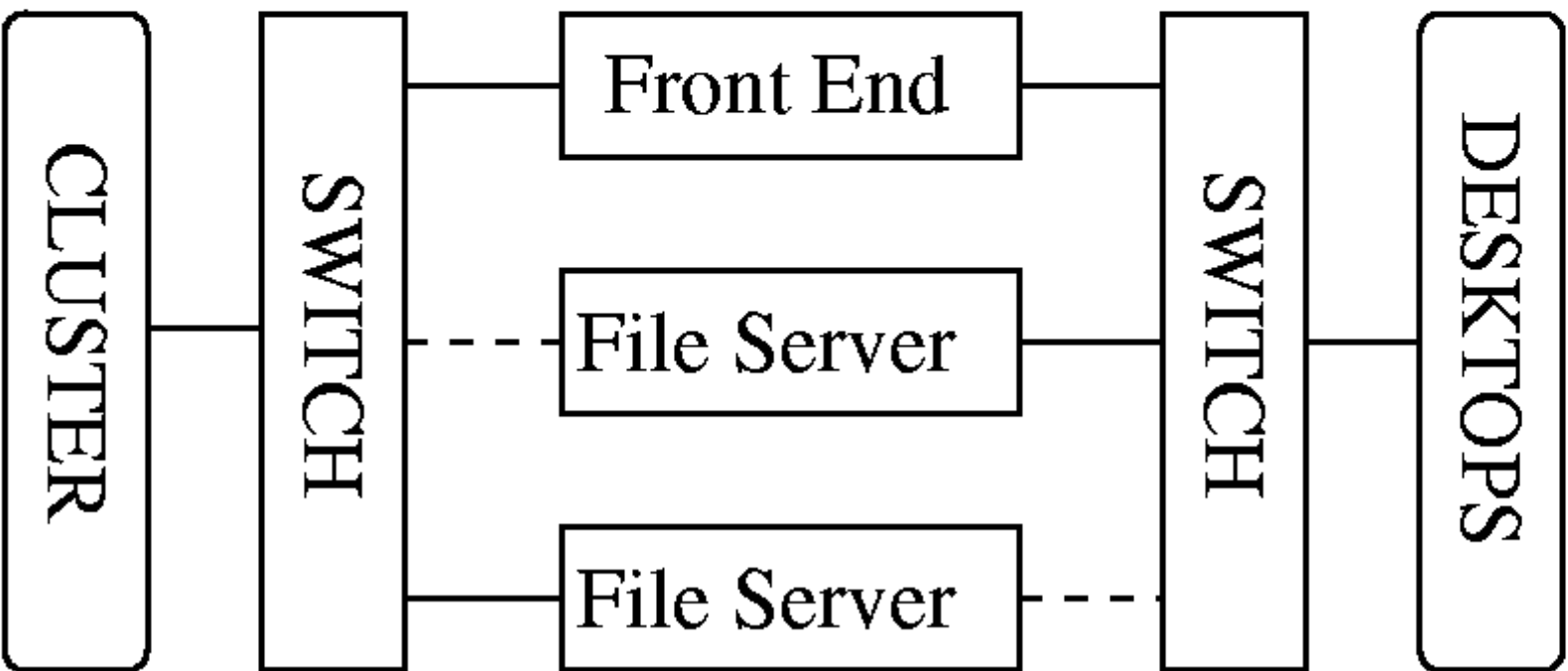
- 445.6 wing, 100k points in grid
- each time history requires 3 hrs on 1 proc
- to map out flutter boundary
  - 7 Mach numbers x 5 calcs/Mach number x 3 hrs
  - total 105 CPU hrs

# Conclusions

- 1997 hardware costing about 40K allowed
  - systematic 2d studies of unsteady aerodynamic phenomena eg aerospike, underexpanded jet, cavity
  - limited 3d investigations of code performance
- 2000 hardware costing about 40K allows
  - systematic parametric studies of flutter
  - detailed evaluation of 3d mechanisms for delta wings and surge waves.....

# What does Beowulf need?

- Relies on cheap yet powerful hardware
- Mature OS software tools
- Must be reliable once assembled
- Homogeneous cluster?
- The first Beowulf was built in 1994
- 1997 the CFD group built their first cluster



# PC Cluster

- New compute nodes 32
  - 750MHz AMD Athlon processors
  - 768 MB of 100 MHz DRAM
  - 100 Mbps duplex 3Com network cards
- Old computer nodes 15
  - 200 PPro processors (Dual)
  - 256 MB of SDRAM
  - 100 Mbps duplex 3Com network cards
- Networking
  - 48 Port Cisco fast Ethernet switch
  - 2 slots for gigabit connections to server

# Need for supporting interface

- Access to different file stores
- Archiving data method
- Post processing data
- Method of accessing the cluster
- Maximizing the performance

# User Driven Requirements

- Glasgow's CFD group use one code for over 90% of all work
- How is the cluster to be used?
- What policies need to be implemented?
- Performance vs Stability

# Software to make it all work

- Some for of message passing
  - PVM 3.4.3
  - MPICH 1.2.2
  - LAM/MPI 6.5.4
- Compiler
  - GCC PGCC Intel's Linux compiler
- Queue system
  - PBS LSF etc.

# FPU on different Nodes

- Only meaningful benchmark is your code base
- Compiled with standard optimizations
- K7 are no faster than a PII 450 (PCFD99)

<b>Operation</b>	<b>PPro 200</b>	<b>K7 750</b>	<b>K7 1000</b>	<b>P2 450</b>
<b>Mat. Vec</b>	46	124	126	115
<b>SAXPY</b>	21	77	80	58
<b>L2 NORM</b>	45	115	129	108
<b>Inner Prod</b>	25	86	99	N/A

# Compiler Performance

- Only meaningful benchmark is your code base
- The GCC results are normalized to 1
  - (1) -O2
  - (2) -O2 -fast -tp athlon -Msafepr
  - (3) -O2 -fast -tp athlon -Msafepr - prefetch

<b>Operation</b>	<b>PGCC<sub>(1)</sub></b>	<b>PGCC<sub>(2)</sub></b>	<b>PGCC<sub>(3)</sub></b>
<b>Mat. Vec</b>	1.25	1.36	1.27
<b>SAXPY</b>	0.87	0.92	0.85
<b>L2 NORM</b>	1.06	1.06	1.06
<b>Inner Prod</b>	1.01	1.02	0.93

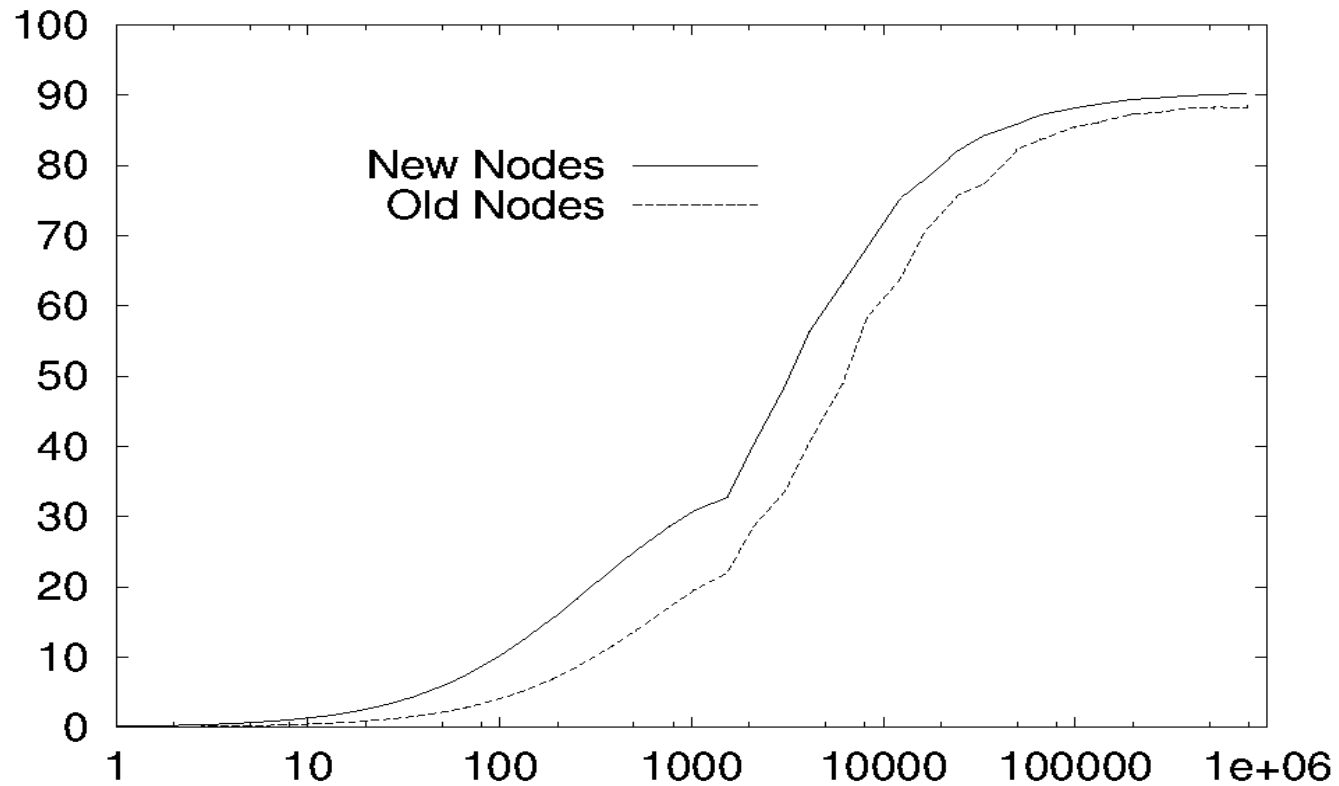
# Network Latency

- Code depended if this is important
- Good algorithm selection/ordering helps
- Very network driver dependent

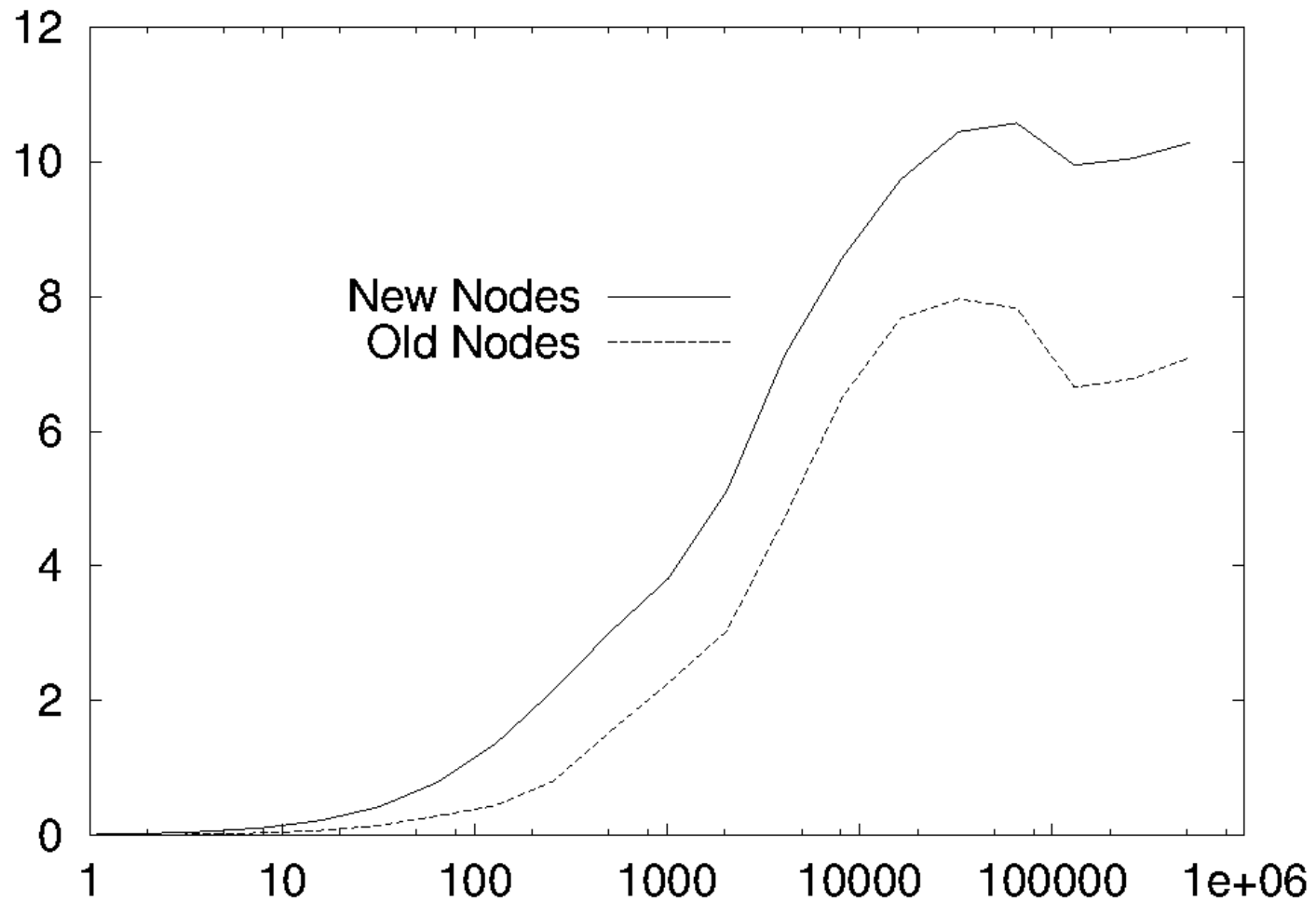
	<b>Old Nodes</b>	<b>New Nodes</b>
<b>TCP</b>	164	56
<b>MPI/LAM</b>	184	63

# Network Bandwidth

- Depends totally on transport layer used



# Network Bandwidth Continued



# I/O Performance (old)

- Server is a bottleneck for some codes
- Clustering the local disks on the nodes?
- The first relate to ASCII then Binary

<b>Procs</b>	<b>Single</b>	<b>Multiple</b>	<b>Single</b>	<b>Multiple</b>
<b>1</b>	86s	86s	10.6s	10.6s
<b>2</b>	86s	42.5s	13.2s	11.2s
<b>4</b>	88s	21.9s	18.2s	12.9s
<b>8</b>	98s	12.5s	38.3s	8.0s

# Administration and Tuning

- Out-of-the-box setup has been VERY stable
- Tuning needed on the Bottlenecks
  - NFS v3 used
  - Updated 3Com drivers required
  - Minimum installs
- If its not broken don't fix it.
  - OS upgrades require tinkering?
  - 3Com drivers always seem to require work!

# Size of Clusters

- What type of jobs will be run?
- How much support is there?
- How fault tolerant is the code base?
- Do you have enough cooling?

# Future of PC's

- What is the next Generation of CPU for?
  - We have NO input.
- Improved PC architecture
  - A new memory bus is needed
    - Nvidia Nforce 420?
    - Performance of i850 will RAMBUS survive?
- Integration of NI's into chipsets/CPU's

# Higher Speed Networks

- Gigabit Ethernet
- Cluster Adapters
  - Myricom's Myrinet
  - Dolphin Interconnects PCI-SCI Cards
  - Giganet cLan
- Light weight Transport layers
  - Gamma
  - M-VIA

# Mass storage and Backup

- How do you provide user space?
  - Is NFS the only solution?
- How do you provide scratch space?
  - PVFS
  - ??
- How do you do backups?
  - Local partial data backups?

# Return of the CoWs?

- Workstations have good FP performance
- Other costs need to be reduced
- PC clusters will always be the cheapest
- Cost vs Performance is code dependent
- Turn around times of the codes
  - As can be seen from PMB we seem to have reached maximum speed on pc's

# Conclusions

- Beowulf PC clusters are cheap.
- Stability is excellent; if not cutting edge.
- Can be tailored to a small groups requirements.
- Allow us to work on problems which would be beyond our computing power.