Using an Enterprise Grid for Execution of MPI Parallel Applications – A Case Study

Adam K. L. Wong and Andrzej M. Goscinski
{aklwong, ang}@deakin.edu.au

School of Engineering and Information Technology
Deakin University
Geelong, Vic 3216, Australia

Enterprise Grids and Parallel Computing on Enterprise Grids

- Many parallel programs can benefit from running even on a small cluster
- An Enterprise Grid is a cluster of clusters
  - If an enterprise owns many geographically dispersed clusters, it could be efficient to connect them together via a fast network to form a cluster of clusters
- Communication latency and the heterogeneous nature of enterprise grids make high performance executions of parallel programs a challenging task

Heterogeneity of an Enterprise Grid

- Cluster heterogeneity
  - Intra-cluster heterogeneity – individual clusters are made up of heterogeneous computers
  - Inter-cluster heterogeneity – if different clusters are internally homogeneous they could be different
- Network heterogeneity
  - A LAN connecting computers in one cluster may be different from that of another cluster
  - The clusters of the enterprise grid are usually connected by a slower WAN – communication bottlenecks usually occur at the inter-cluster communications

Work Objective

Aim:

- Carry out a study into the execution performance of MPI parallel applications on two enterprise grids

Tasks:

- To study whether it is feasible to employ enterprise grids to carry out high-performance parallel computation – **co-allocation** of computers from different clusters
- To show how to improve the execution performance of parallel programs on an enterprise grid using dynamic load balancing based on the source initiative strategy

MPI Parallel Applications – Selection Criteria

- Applications must be well known
  - We used both benchmarks and real applications
- They must satisfy the following attributes:
  - *Computation attributes*
    - Problem size of a parallel program
    - Computation Bound vs Communication Bound
  - *Communication attributes*
    - Communication Volume and Communication Pattern
  - *Memory attributes*
    - Large but such that memory swapping is avoided
  - *Topology attributes*
    - Process-to-Processor Mapping: size and structure

MPI Parallel Applications – Classification of the Selected Programs

MPI-Povray
Embarrassingly parallel (EP)
Block Tridiagonal Solver (BT)

Parallel fastDNAml
LU Solver (LU)
Multigrid (MG)

Program	Selection Attributes			
	Computation	**Comm. Volume**	**Comm. Pattern**	**Topology**
MPI-Povray	Comput. Bound	Low	Point-to-Point	Any
Parallel fastDNAml	Comput. Bound	Low	Point-to-Point	Any
EP	Comput. Bound	Negligible	Point-to-Point	Any
LU	Communic. Bound	Low	Point-to-Point	Power-of-2
BT	Communic. Bound	Medium	Collective	Square-of-n
MG	Communic. Bound	High	Collective	Power-of-2

Experiments – Execution Measurements

- We studied the influence of two dimensions of heterogeneity in enterprise grids: (1) inter-cluster network and (2) inter-cluster computers on the execution performance of parallel applications
- In the experiments, the processes of a parallel application were co-allocated (one process per computer) on the two DEG clusters to gain a higher level of parallelism and by this to hopefully achieve better execution performance
- Clusters CG_n (Geelong) and CM_n (Melbourne) had n computers in each cluster

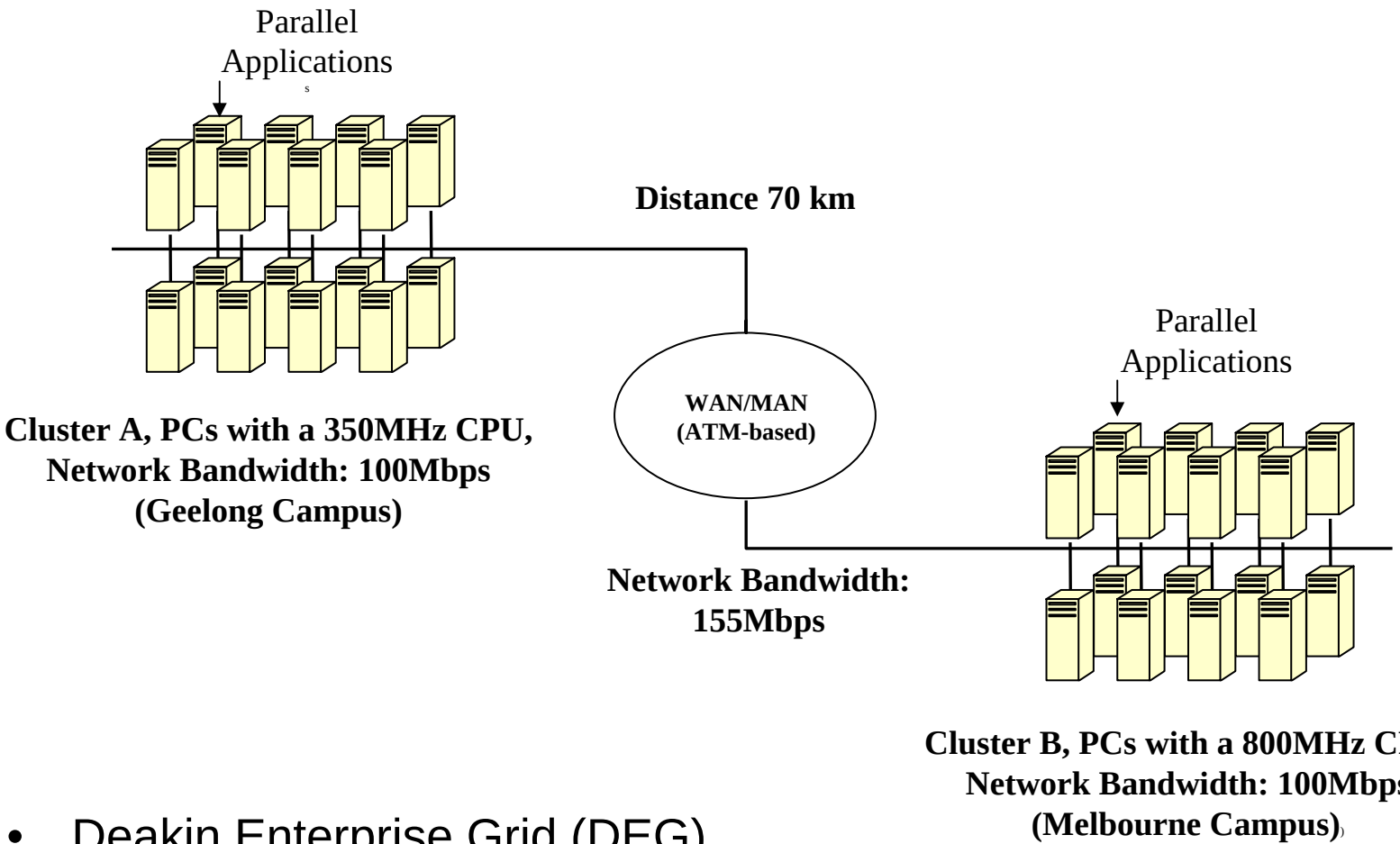
Experiments – Execution of the NAS Parallel Benchmarks

- Experiment 1: 16 processes of a program co-allocated on 16 computers of two clusters
- Experiment 2: different numbers of processes of a program co-allocated in equal numbers on computers of two clusters
- The objectives were to
  - Study the feasibility of high-performance computing on enterprise grids
  - Study the influence of grid networks' bandwidth, and program's computation and communication characteristics on its execution performance

Experiments – Execution of the NAS Parallel Benchmarks

- All experiments were carried using EP, LU, BT and MG
- Some important issues:
  - Grid networks bandwidth of
    - LAN and WAN/MAN comparable
    - LAN and WAN/MAN different
  - Program Size:
    - Class B of the NAS parallel programs
  - Program Topology:
    - As specified
  - Execution Time of the Programs:
    - The subject of our study

Enterprise Grid 1



- Deakin Enterprise Grid (DEG)

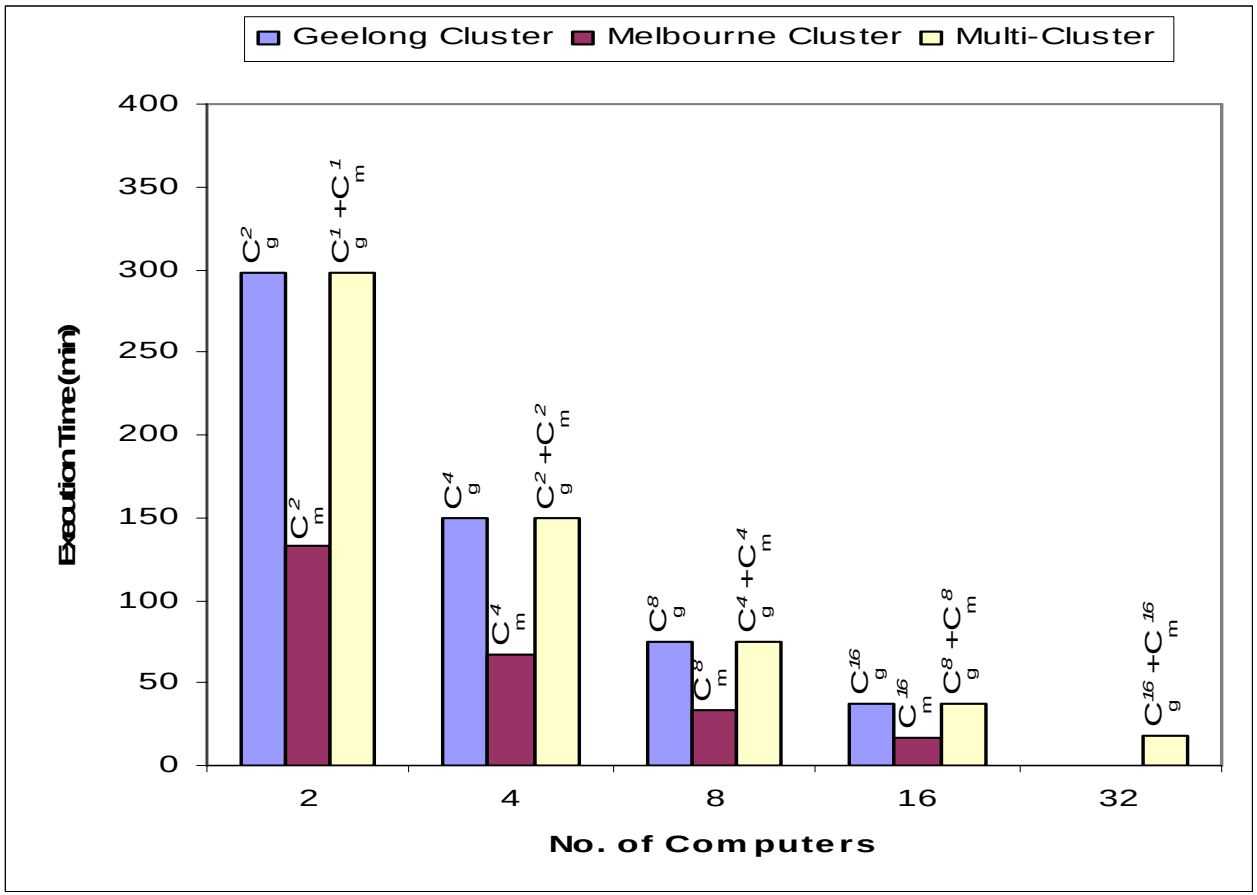
Execution of NAS Parallel Benchmarks (Result of Experiment 1)

- Executions of the NAS Programs with different co-allocation ratios on two clusters

Co-allocation Ratio		Execution Time of NAS Programs (min)			
No. of Computers in C_G^n	**No. of Computers in C_M^{16-n}**	**EP**	**LU**	**BT**	**MG**
16	0	37.3	41.8	47.1	37.9
14	2	37.4	41.9	48.9	45.5
12	4	37.3	42.2	59.8	47.9
10	6	37.4	42.4	58.7	53.9
8	8	37.7	44.5	61.3	61.0
6	10	37.3	45.2	61.7	47.7
4	12	37.8	41.8	62.3	57.7
2	14	37.2	40.7	49.6	43.4
0	16	16.6	27.9	33.3	30.0

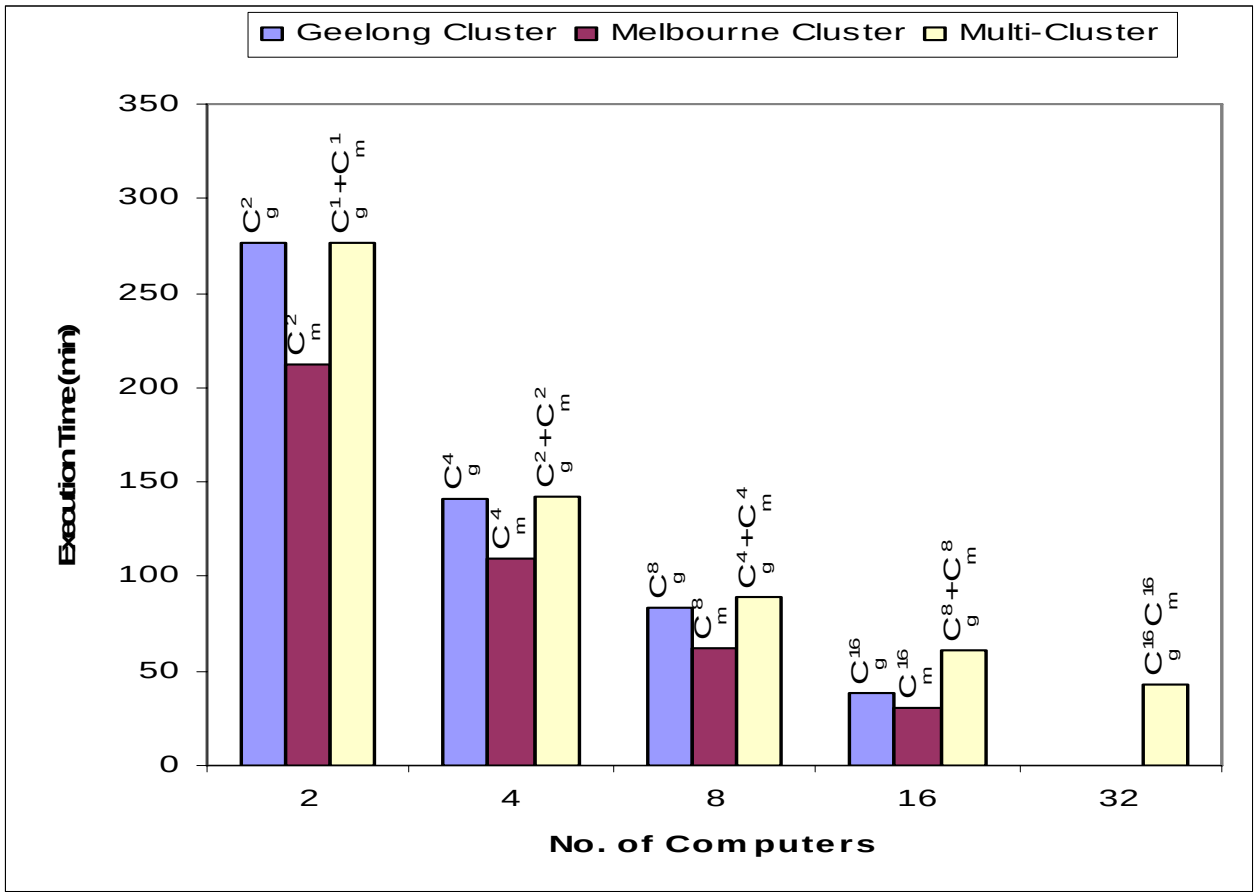
Execution of EP (Result of Experiment 2)

- Improvement despite of the inter-cluster communication cost
- Results of LU and BT are very similar



Execution of MG (Result of Experiment 2)

- Improvement despite of the inter-cluster communication cost



Experiments – Execution of the MPI-Povray and Parallel fastDNAml

- We carried out the same two sets of experiments (1) and (2) using two real parallel applications: MPI-Povray and Parallel fastDNAml
- The objectives were to
  - Support our claim that high-performance computing can be practically conducted on enterprise grids
  - A simple but effective program-level dynamic load balancing approach adopted in these two applications was studied to provide better execution performance of the programs

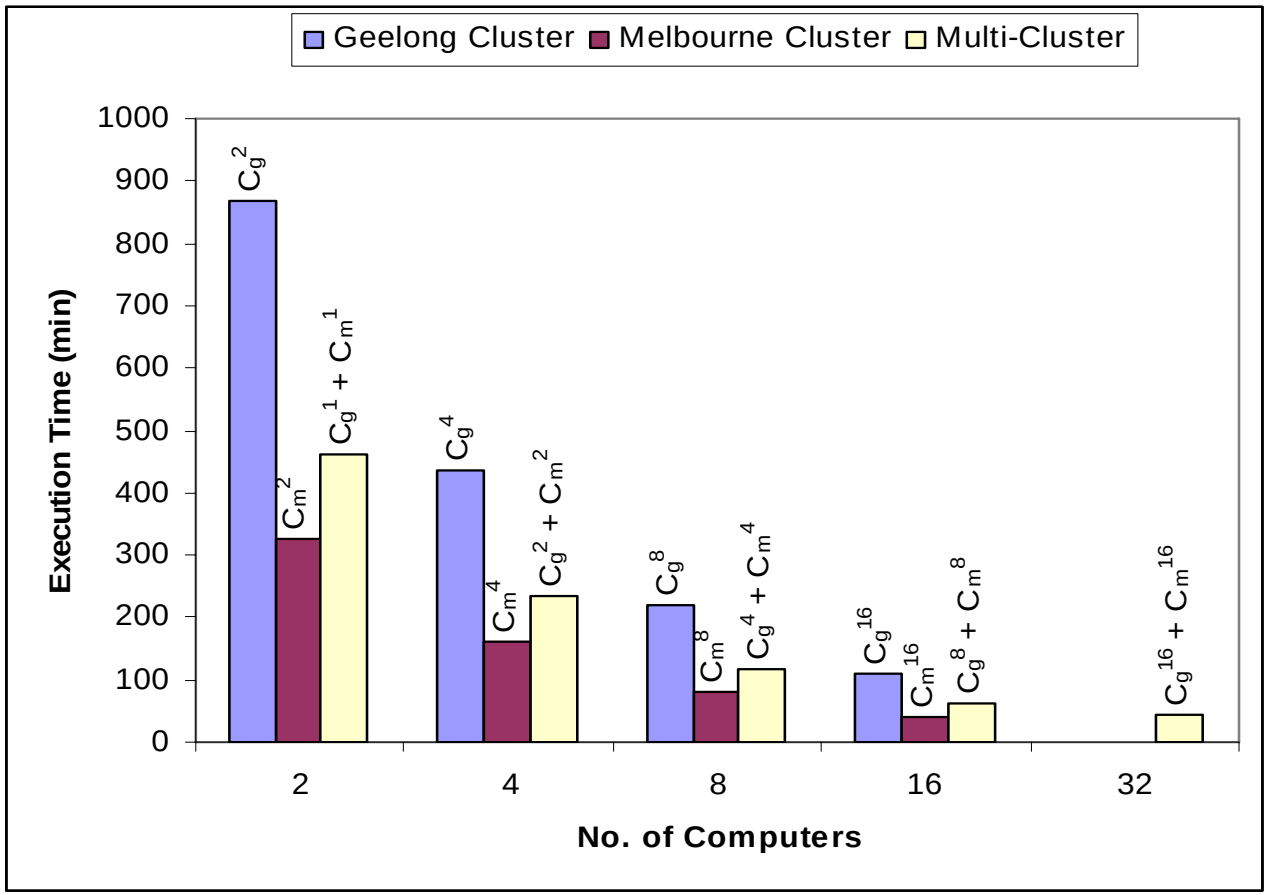
Execution of the MPI-Povray and Parallel fastDNAAml (Result of Experiment 1)

- Execution of the MPI-Povray and Parallel fastDNAAml with different co-allocation ratios on two clusters

Co-allocation ratio		Execution Time (min)	
No. of Computers in C_g^n	**No. of Computers in C_m^{16-n}**	**fastDNAAml**	**MPI-Povray**
16	0	173.3	110.3
14	2	160.8	92.3
12	4	147.9	77.7
10	6	138.9	69.4
8	8	129.2	61.2
6	10	122.2	54.6
4	12	114.6	50.1
2	14	108.6	45.7
0	16	105.3	41.2

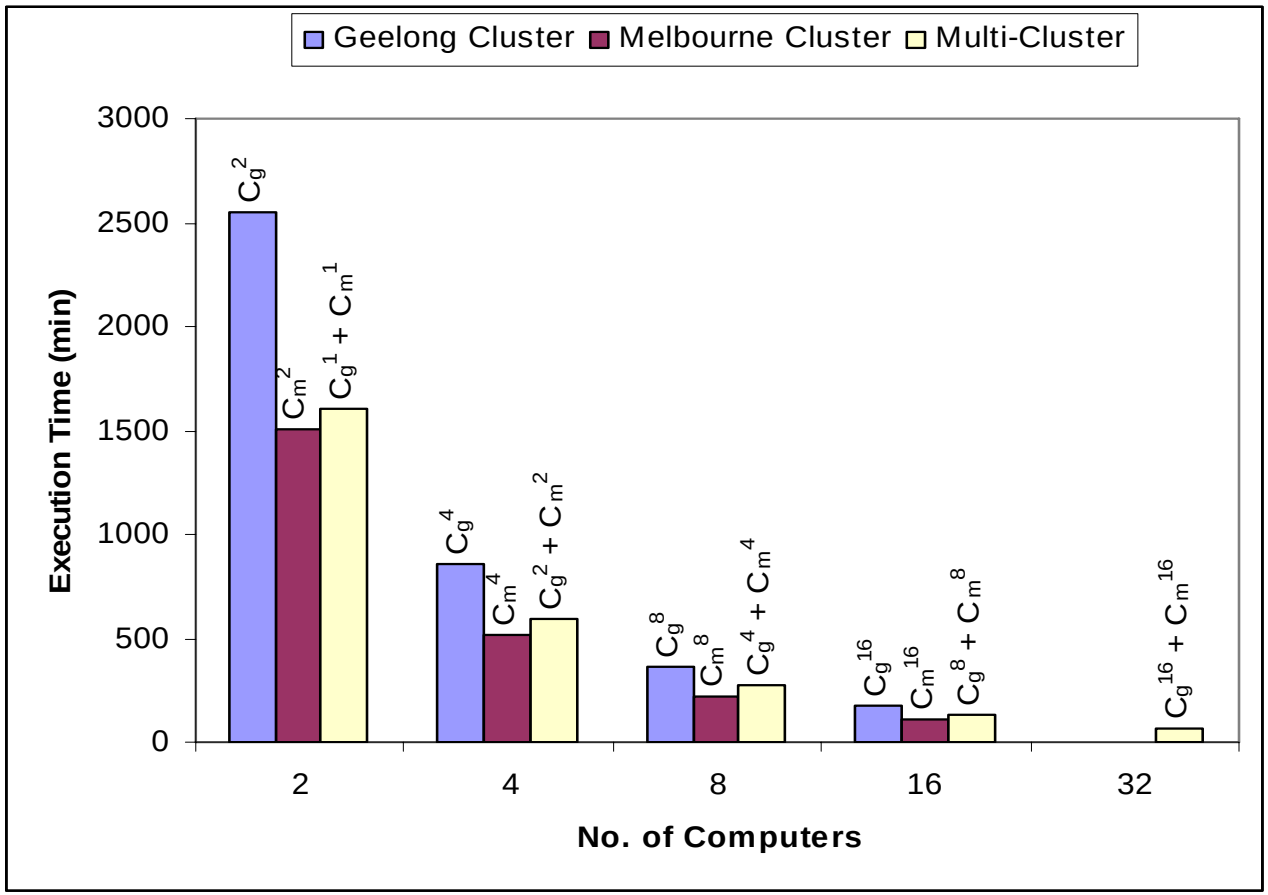
Execution of MPI-Povray (Result of Experiment 2)

- Improvement despite of the inter-cluster communication cost

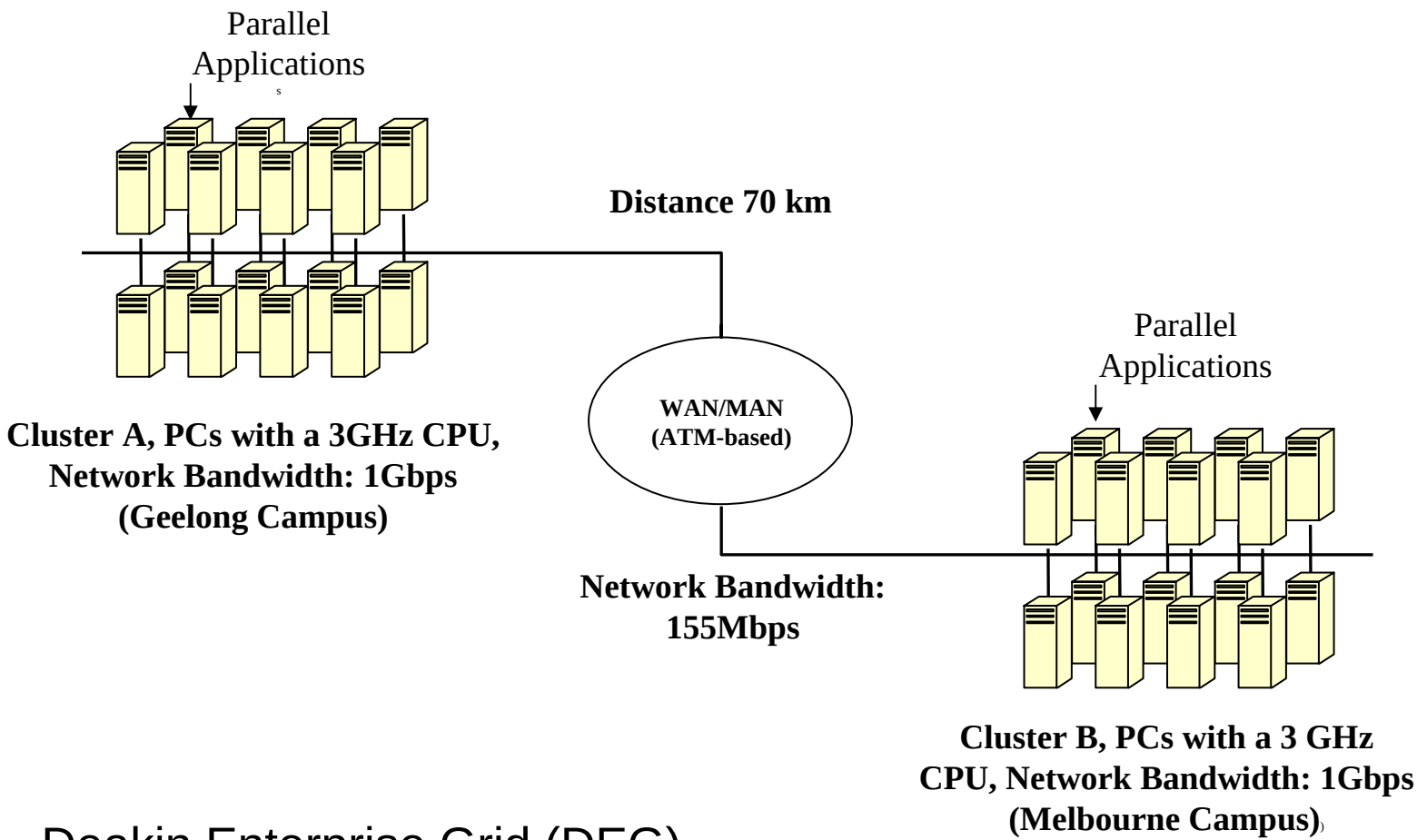


Execution of Parallel fastDNAAm (Result of Experiment 2)

- Improvement of despite of the inter-cluster communication cost



Enterprise Grid 2



- Deakin Enterprise Grid (DEG)

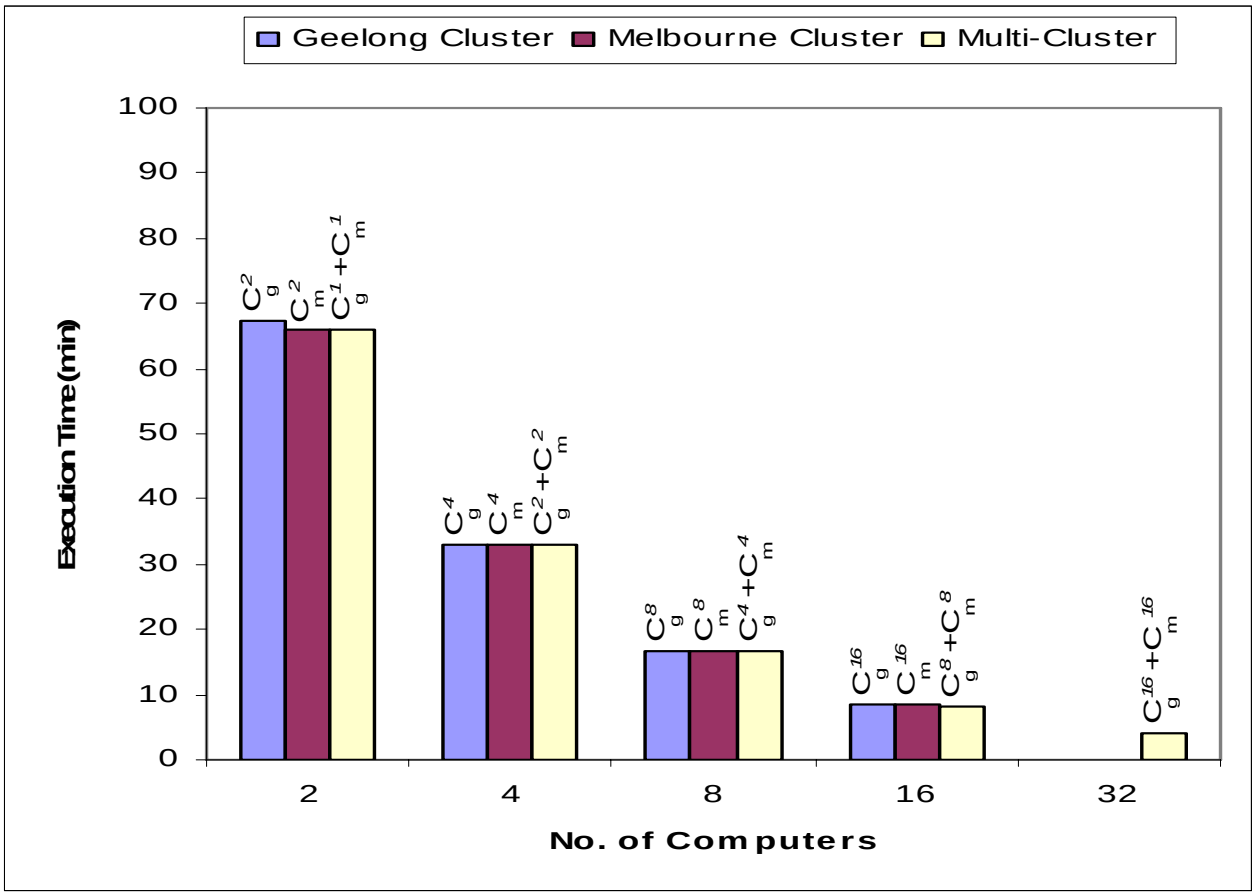
Execution of NAS Parallel Benchmarks on Windows-PC based Clusters (Experiment 1)

- Executions of the NAS Programs with different co-allocation ratios on two clusters

Co-allocation ratio		Execution Time of NAS Programs (min)		
No. of Computers in C_g^n	**No. of Computers in C_m^n**	**EP**	**LU**	**MG**
16	0	8.48	7.14	6.64
14	2	8.32	7.17	29.54
12	4	8.38	9.74	39.22
10	6	8.34	8.25	47.29
8	8	8.31	8.75	44.45
6	10	8.33	8.49	46.20
4	12	8.31	11.79	49.02
2	14	8.32	8.66	39.78
0	16	8.58	8.53	7.53

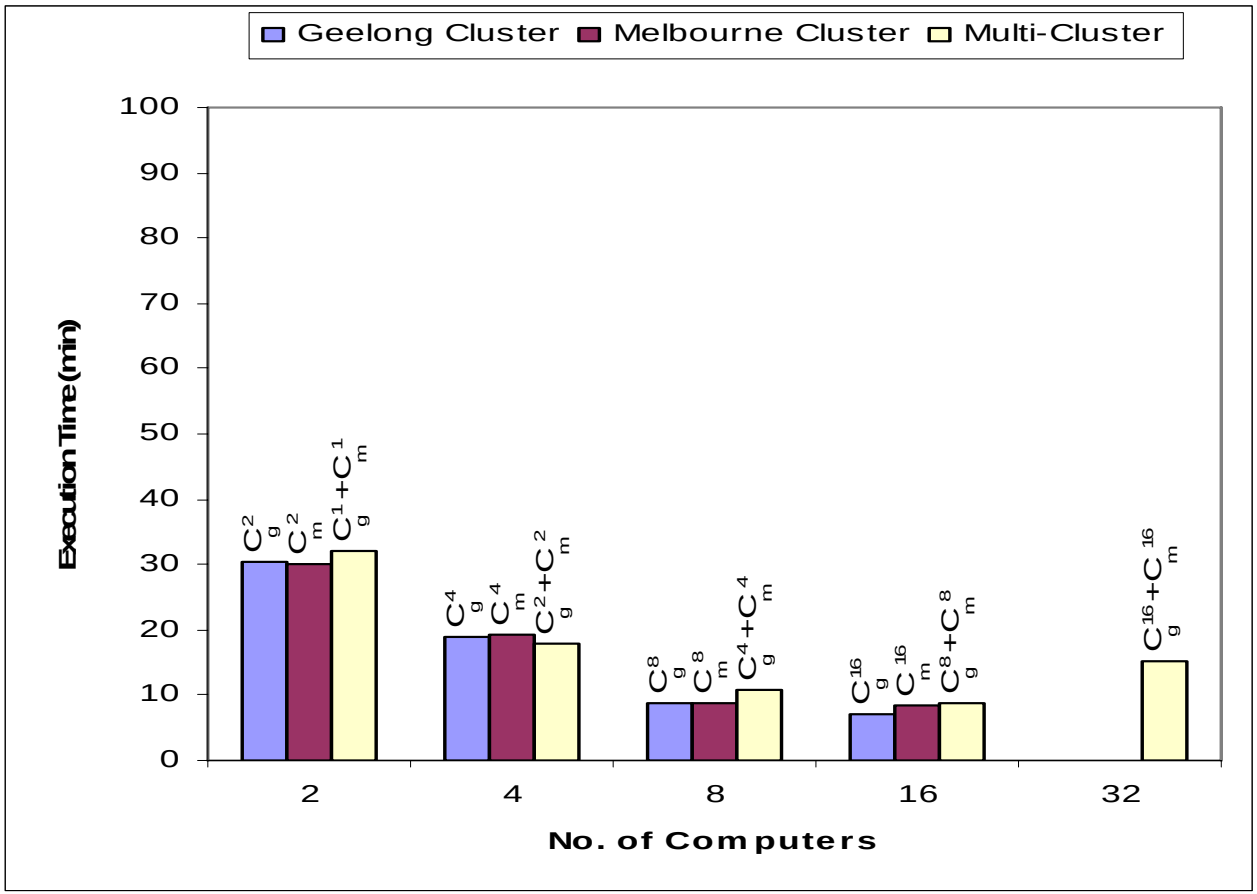
Execution of EP on Windows-PC based Clusters (Experiment 2)

- Positive speedup despite of the inter-cluster communication cost



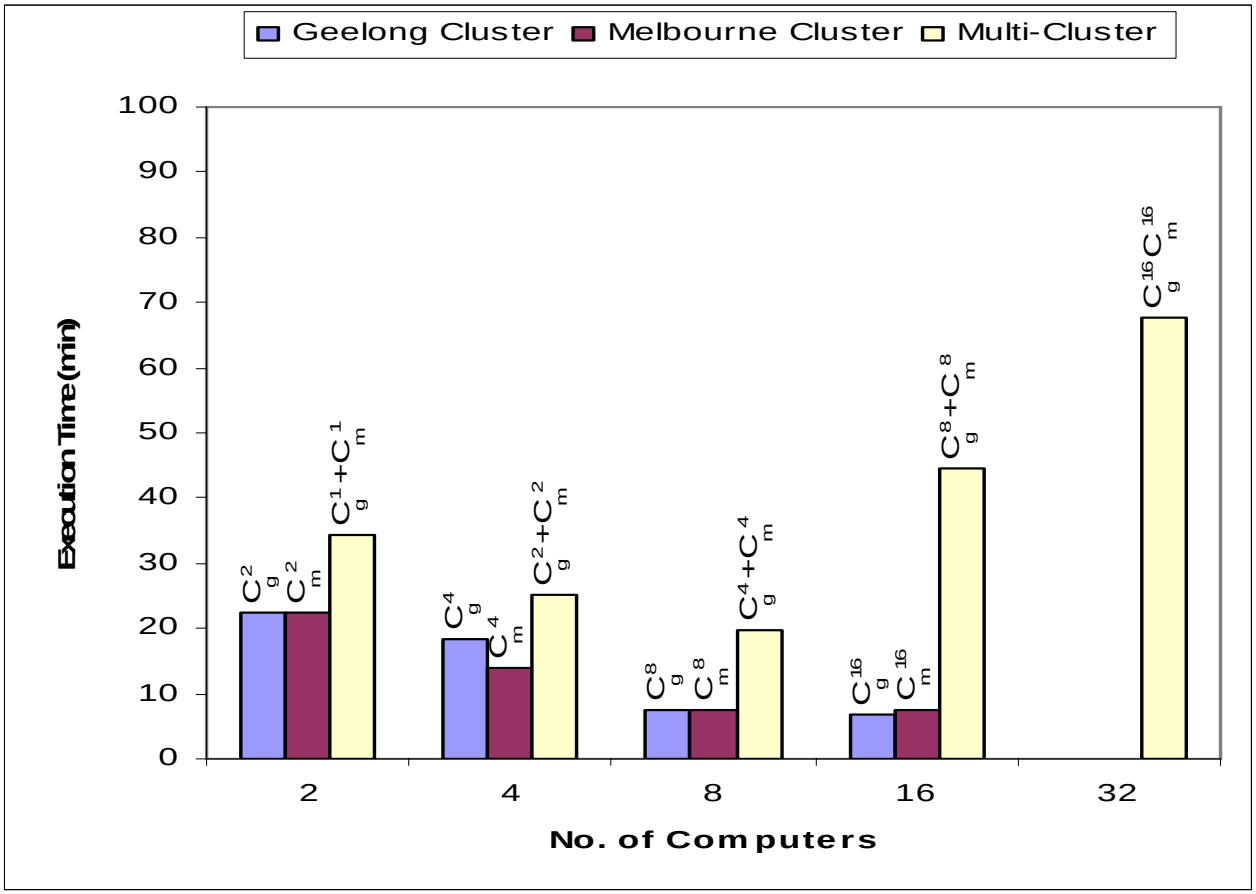
Execution of LU on Windows-PC based Clusters (Experiment 2)

- Positive speedup despite of the inter-cluster communication cost



Execution of MG on Windows-PC based Clusters (Experiment 2)

- Negative speedup due to the inter-cluster communication cost



Conclusions

- The results of the study demonstrate that computer co-allocation on multiple clusters for high performance could only be feasible and sound if bandwidth of LANs and bandwidth Wan/MANs are comparable
- Inter-cluster network heterogeneity – NAS experiments
  - computation-bound: EP and LU run well on the enterprise grid
  - communication-bound: BT and MG can also benefit from increased number of computers and thus improve the execution performance despite the inter-cluster communication cost
- Inter-cluster computers heterogeneity – MPI-Povray and Parallel fastDNAml experiments
  - the execution improvement gained through balancing the workloads on computers of the enterprise grid can outweigh the inter-cluster communication cost