Scalability of Weather & Climate Codes

ECMWF Workshop Nov04’08

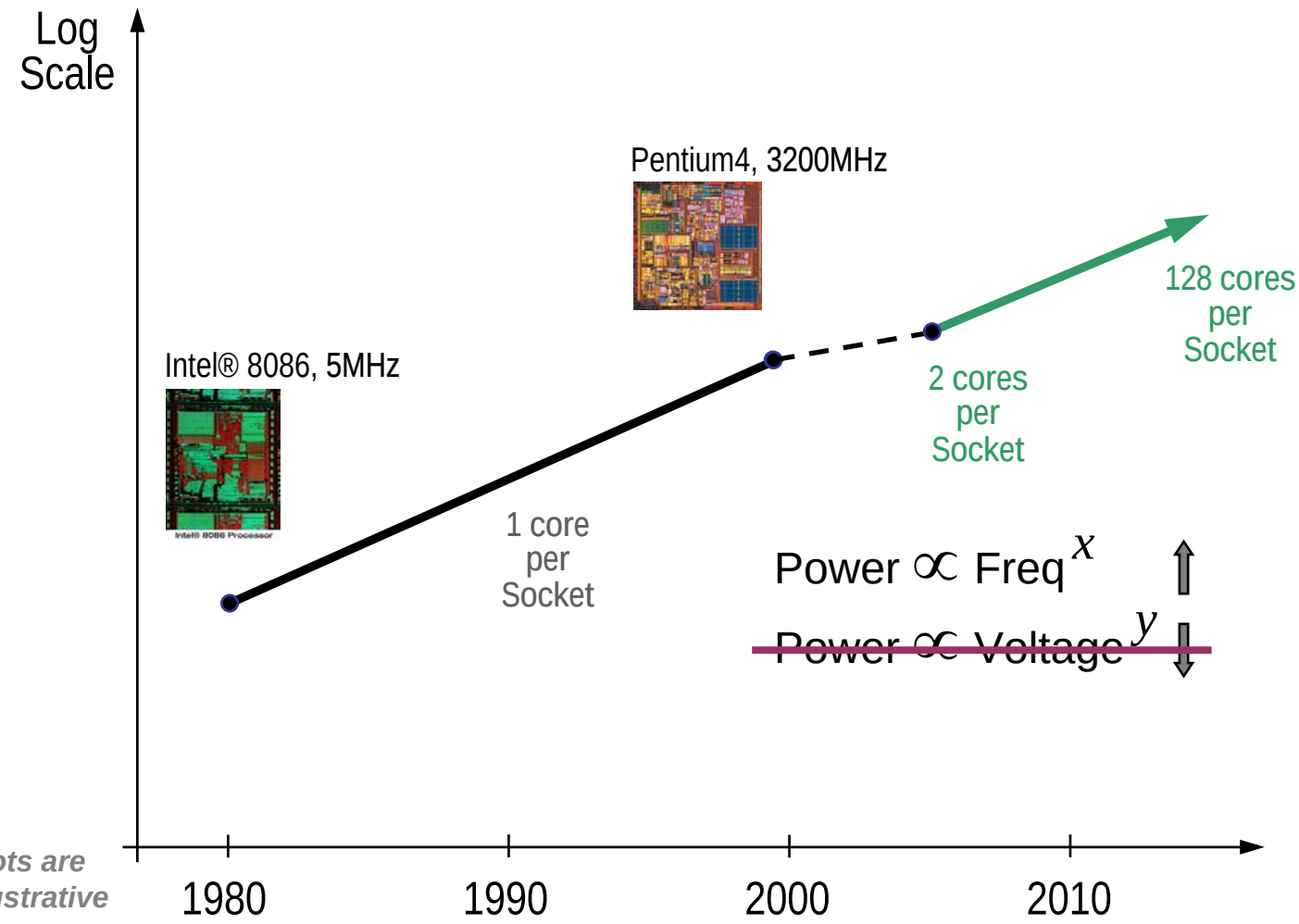
Back to Basics : Proposed New Architecture for Many-Core

- Moore's Law Result Shift : Clock → Many-Core
- Problem Shift : Scaling Software Applications
- Load Balancing + Communications Costs
- New Architecture : Communications Focused

PGAS
+
Programmable
Communications
Accelerators

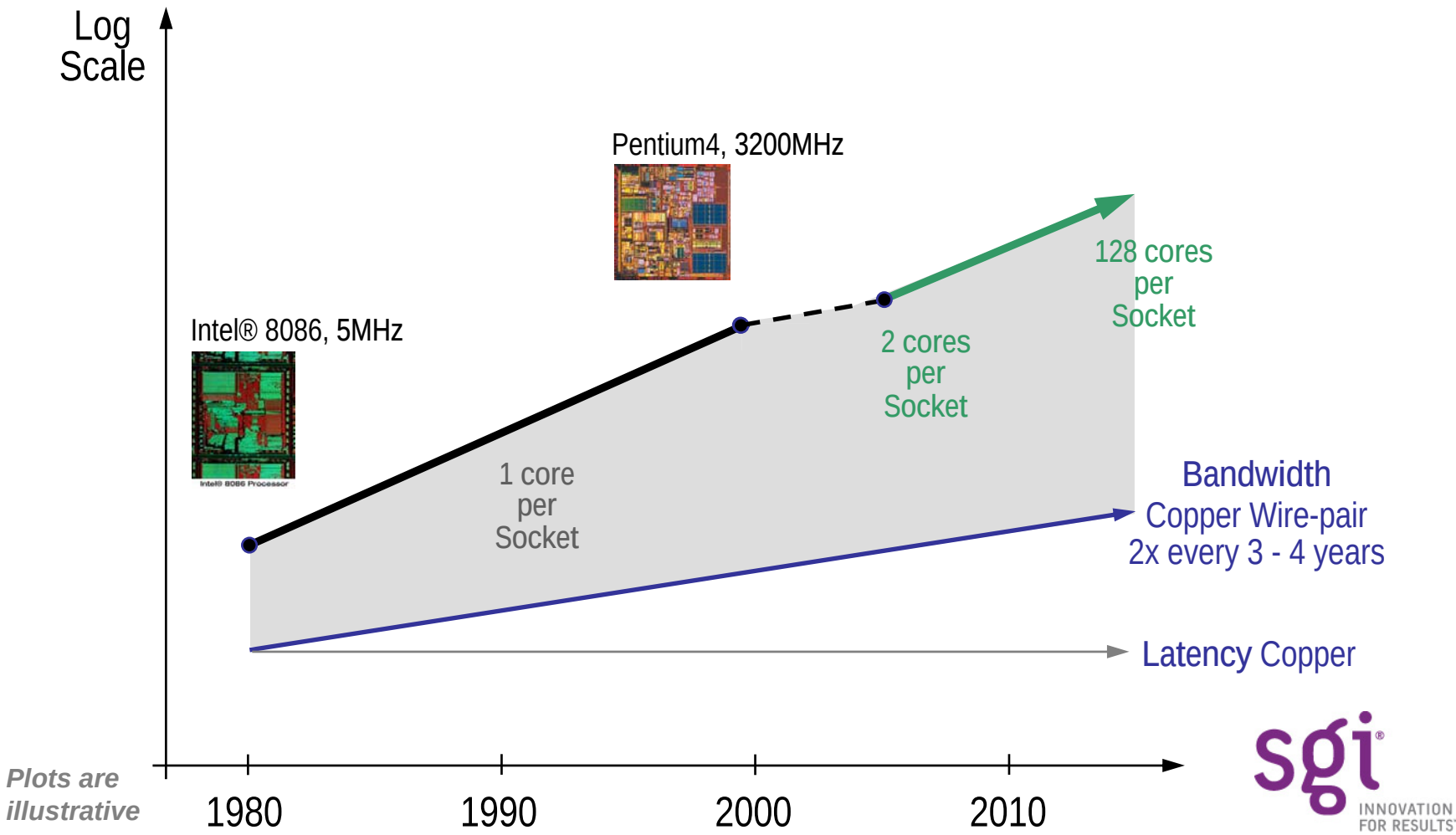
sgi® INNOVATION FOR RESULTS™

Trends

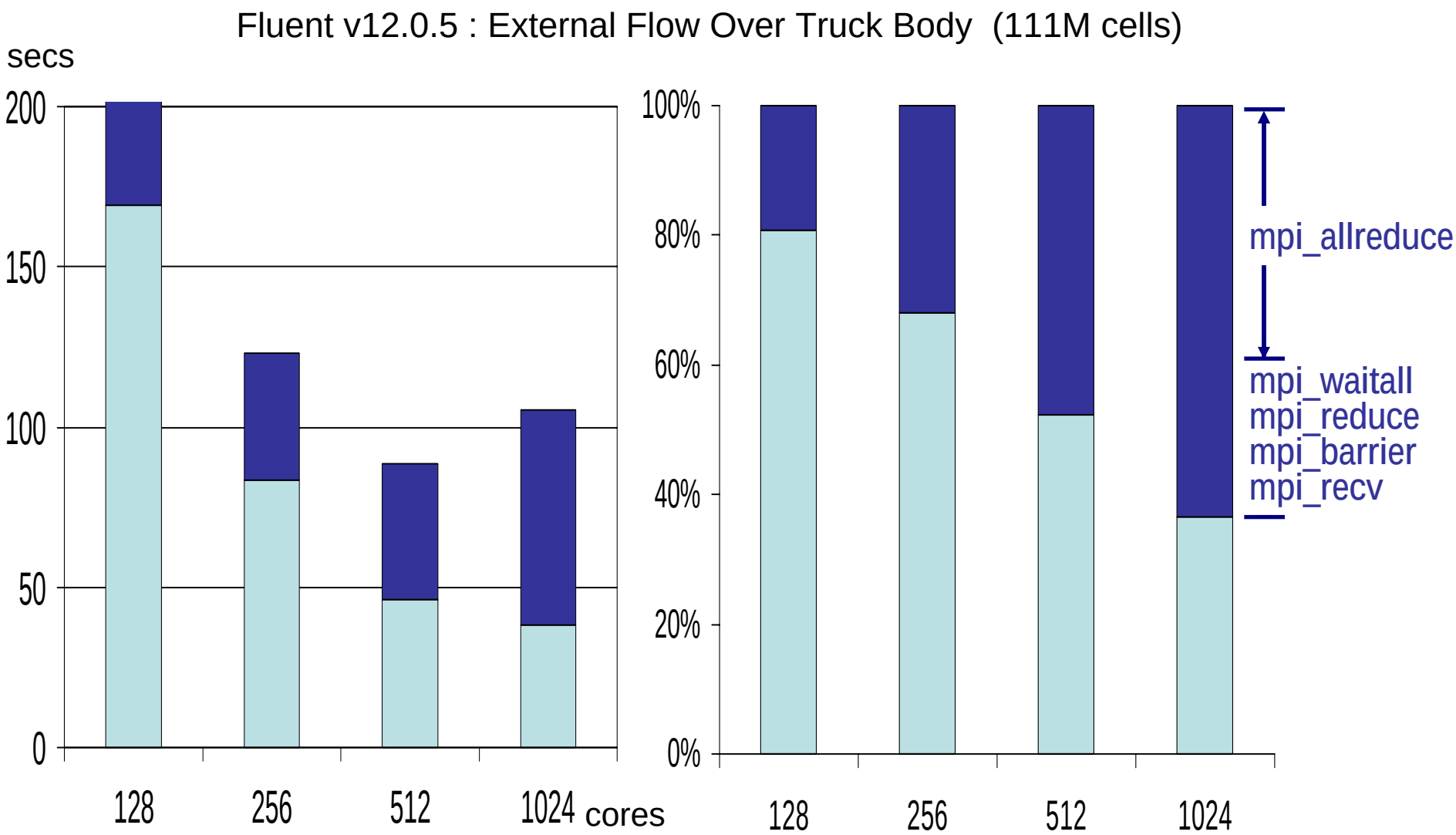


sgi INNOVATION FOR RESULTS

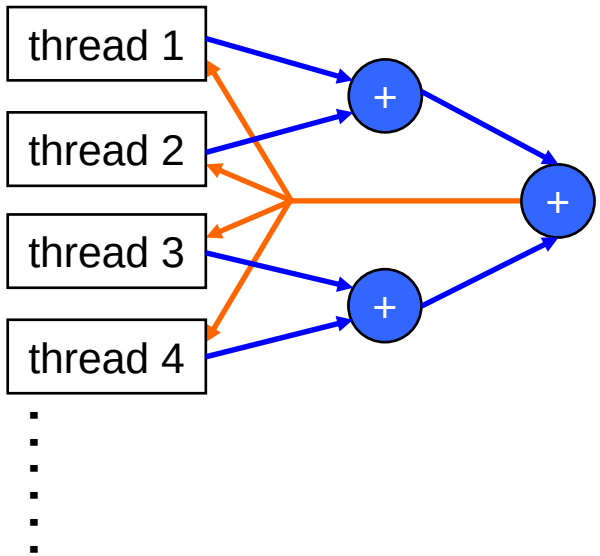
Trends



Total Run Time = Computation + Communications (bw & latency)



Latency : Pay a lot for relatively little gain (e.g. Global Collectives)



mpi_allreduce ();

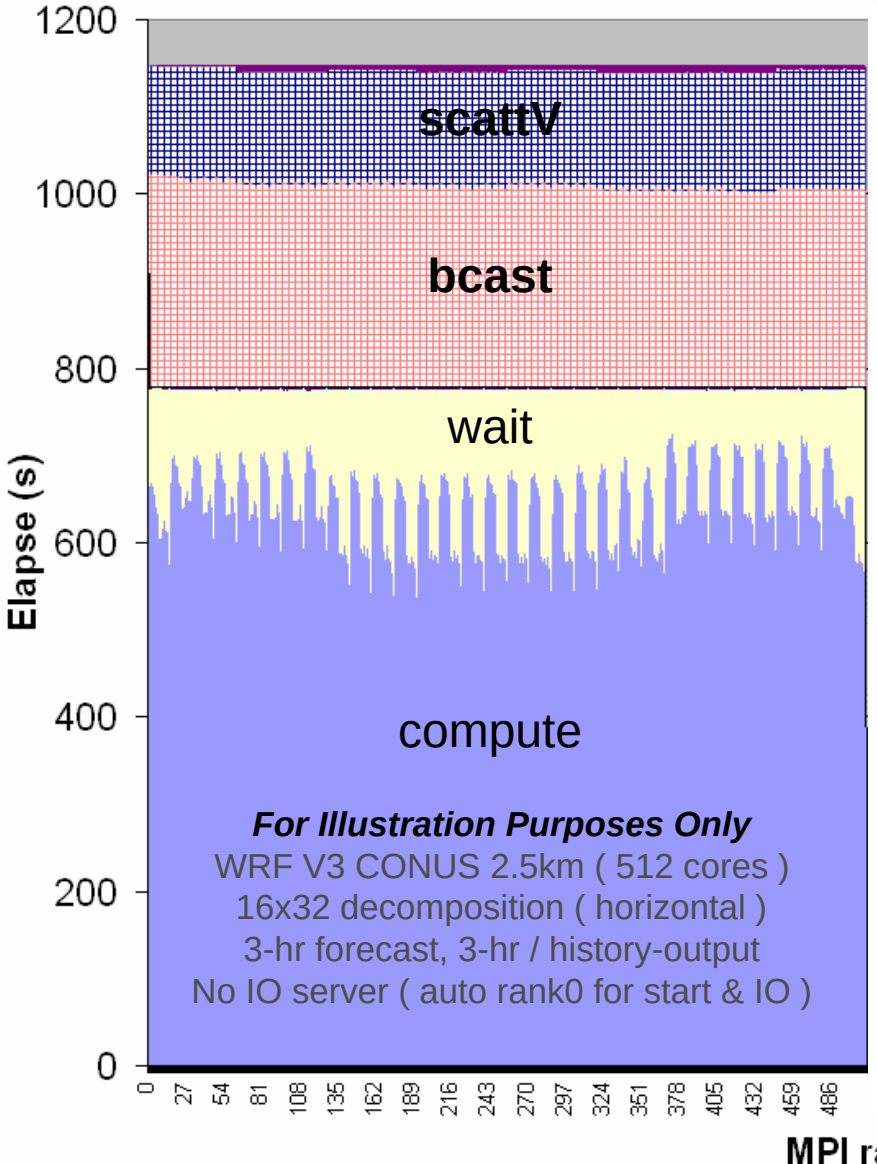
Interconnect

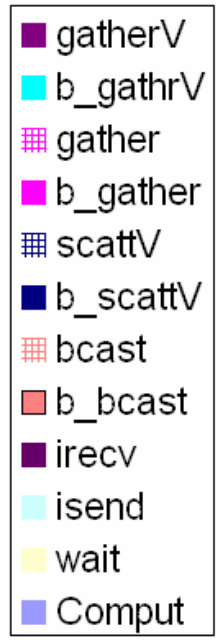
mpi_allreduce ();

```
#pragma omp parallel \
  reduction(operator:list)
upc_all_reduce
```

- GigE (64^3=256K mpi ranks) 100s us
- IB 4x DDR (64^3=256K mpi ranks) 10s us

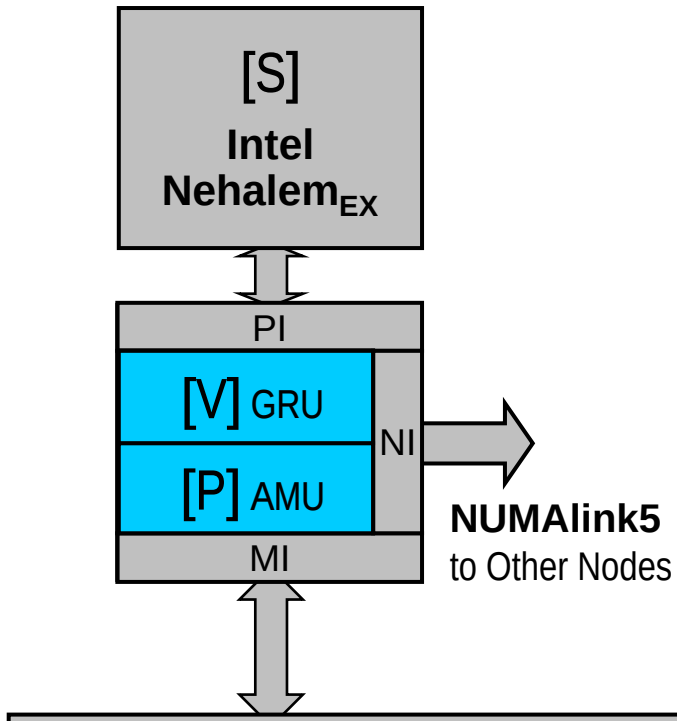
Total Run Time = Computation + Communications (bw & latency)





sgi INNOVATION FOR RESULTS

Ultraviolet : PGAS + PCP Architecture (accelerate communications)



GSM – cc = GAM

GSM

- Partition Memory (OS)
  - Max. 2KC 16TB ('09)

GAM

- PGAS Memory (X-Partition)
- PCP (GRU + AMU)
  - Accelerate PGAS Codes
  - Accelerate MPI Codes (MOE v.v. TOE)

sgi INNOVATION FOR RESULTS

Ultraviolet : PGAS + PCP Architecture for…

PGAS API

- UPC library support
- CAF F95 extensions (formally in Fortran-2008)
- Others welcome suggestions & support

Incumbent API

- MPI-1 library support
- MPI-2 library support
- SHMEM
- OpenMP

sgi® INNOVATION FOR RESULTS™

Back to Basics : Proposed New Architecture for Many-Core

- Moore's Law Result Shift : Clock → Many-Core
- Problem Shift : Scaling Software Applications
- Load Balancing + Communications Costs
- New Architecture : Communications Focused

PGAS
+
Programmable
Communications
Accelerators

Scalable Commodity Cluster
but with all its memory PGAS

sgi INNOVATION FOR RESULTS