# Assignment 1 Report

Corey Crosser

February 10, 2017

CS5463: Cloud and Big Data

Homework 1

Group 9: Lee Boyd, Sean Soderman, Corey Crosser

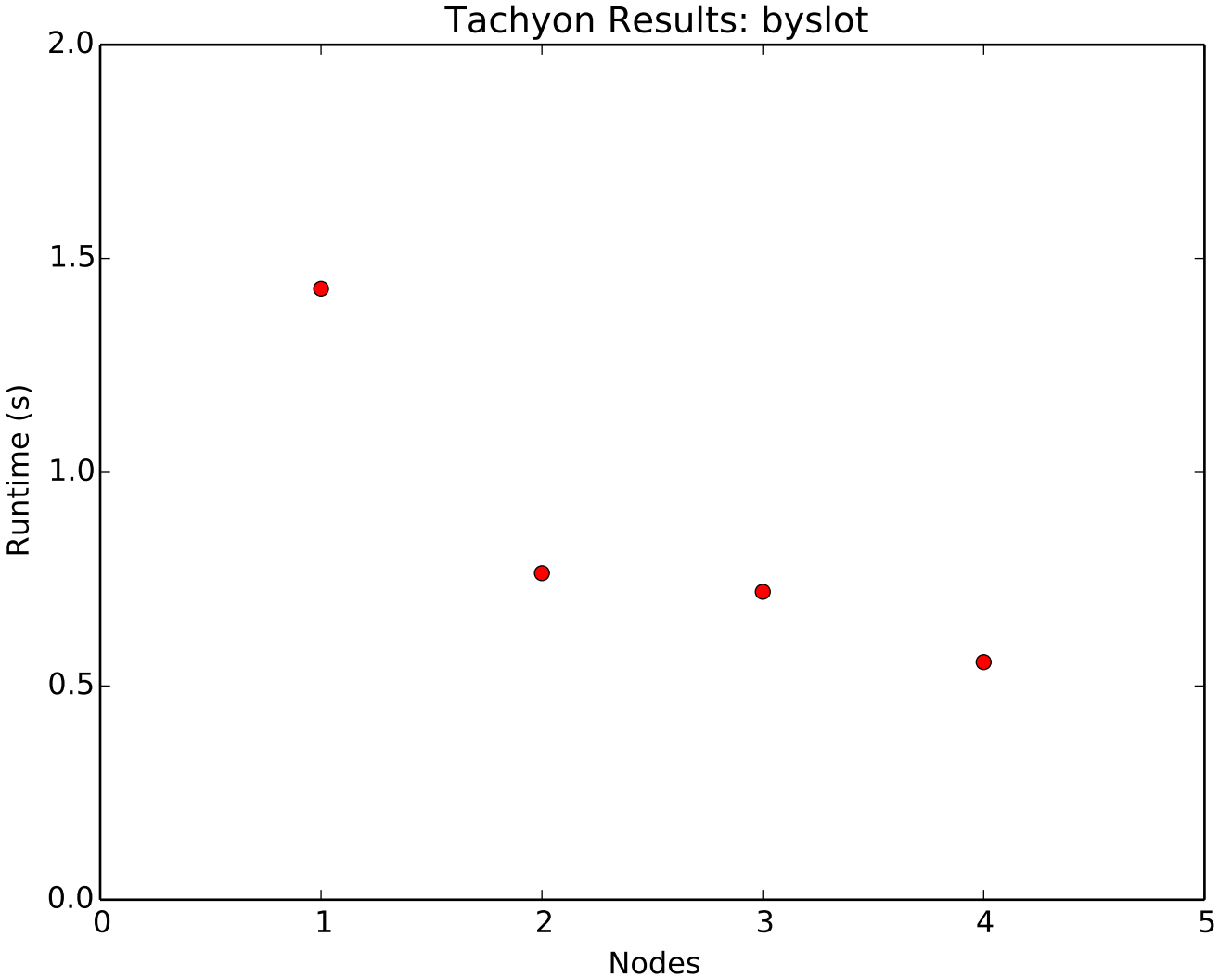
Q1) The results show that by slot and by node scheduling are relatively equivalent and the benchmark results improve logarithmically as machines are added to the cluster. This makes sense because the VMs only have one core, making by slot and by node scheduling roughly equivalent. The sub-linear performance improvement also makes sense because more communication overhead is added with each node, stifling performance gains. This experiment would be more interesting if each node had multiple cores to highlight the difference in these two MPI scheduling policies. See pages 2 and 3 for the first two graphs which illustrate the performance Ray Trace benchmark.

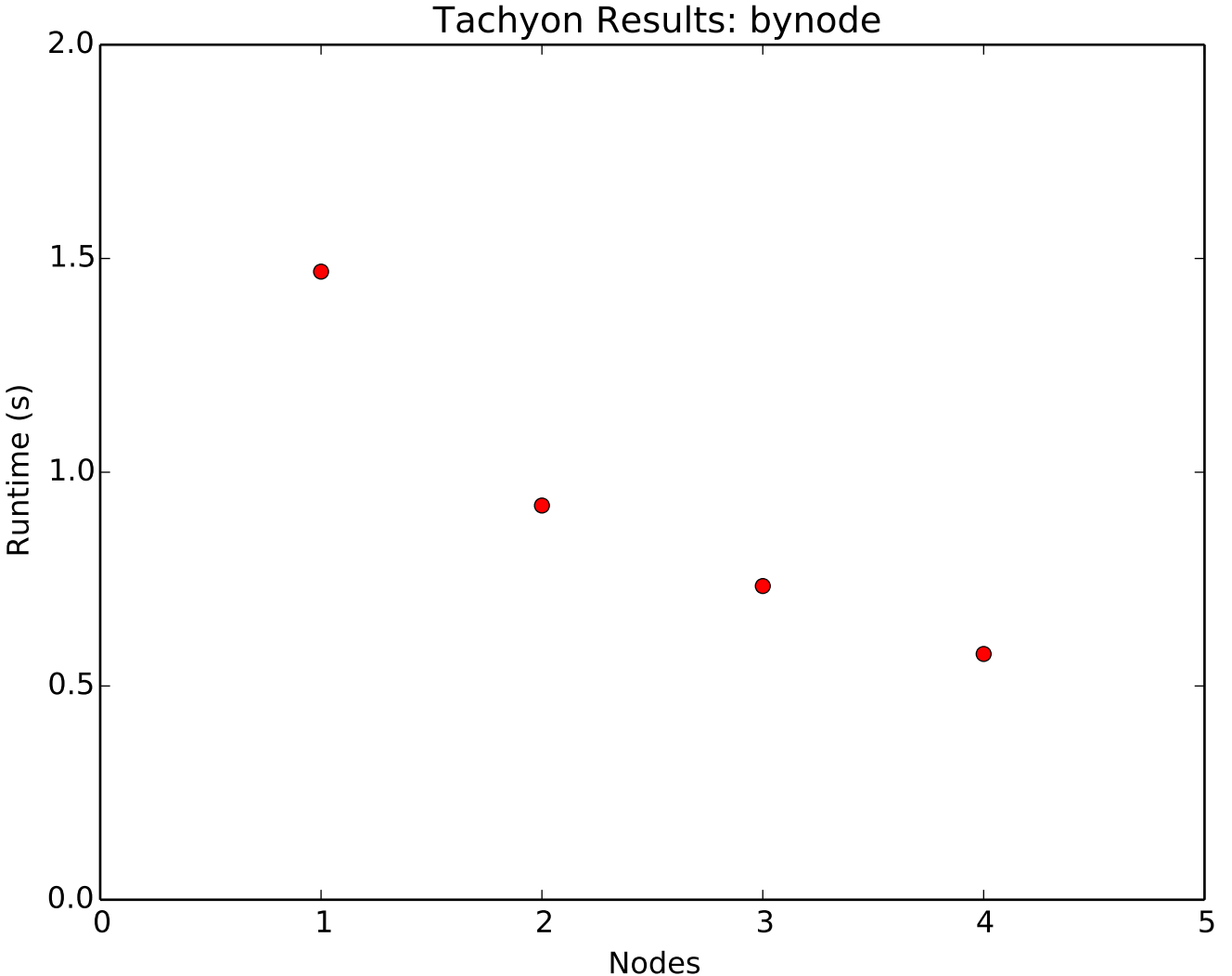
Q2) Our results show an exponential decrease in performance for the PTRANS benchmark as nodes are added. This is counterintuitive to the idea that more nodes should bring more processing power and improve benchmark performance. This result is probably due to our lack of understanding of the HPCC benchmark. We likely did not tune the parameters properly and the provided website was little help. We attempted to tune these numbers manually and the best values we could find for P, Q, N and NB are listed in the table on page 4. These parameters were difficult to determine due to inconsistent results between runs (different performance numbers with same input parameters) and conflicting results between the two HPCC benchmarks (parameters that improve the performance of HPL hurt the performance of PTRANS). The HPL benchmark shows a general improvement as nodes are added to the cluster. As to which MPI scheduling policy is better, that is unclear from our results. See graphs on pages 5-8 for HPCC results.

Q3) Our script works as expected and changes the value of Q to be equal to the number of nodes in the cluster and we chose P=1 for all runs of HPCC. Although we think we may have found some good possible values for N relating to each cluster size, our script does not tune this parameter. We were unable to see any correlation to the NB parameter and HPCC performance. We chose NB=96 for all cluster sizes.

Division of Labor:

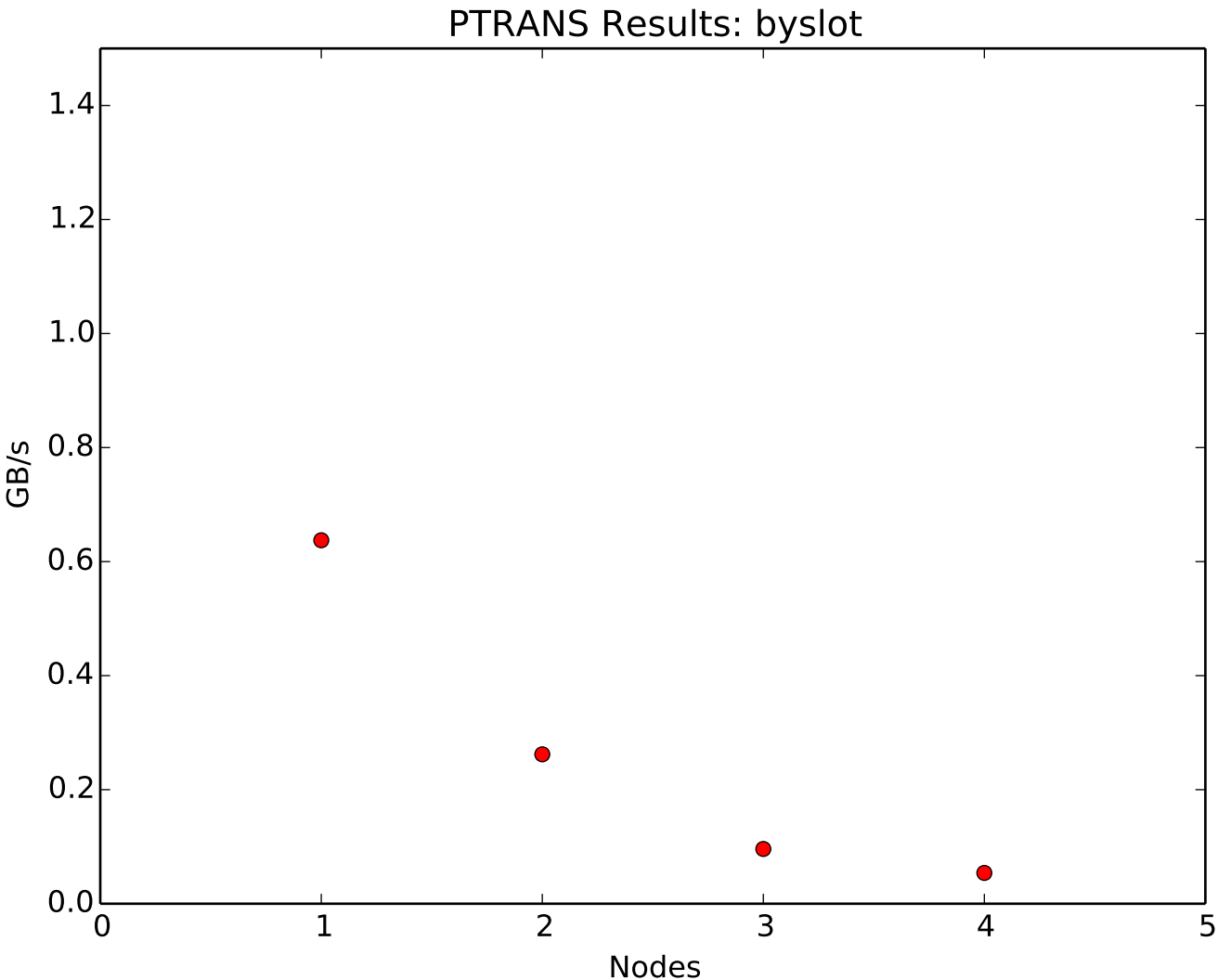
- Cluster Setup: All of us
- MPI Setup: Lee and Sean
- Tuning HPCC Parameters: Corey and Lee
- Python Script: Sean
- Report: Corey

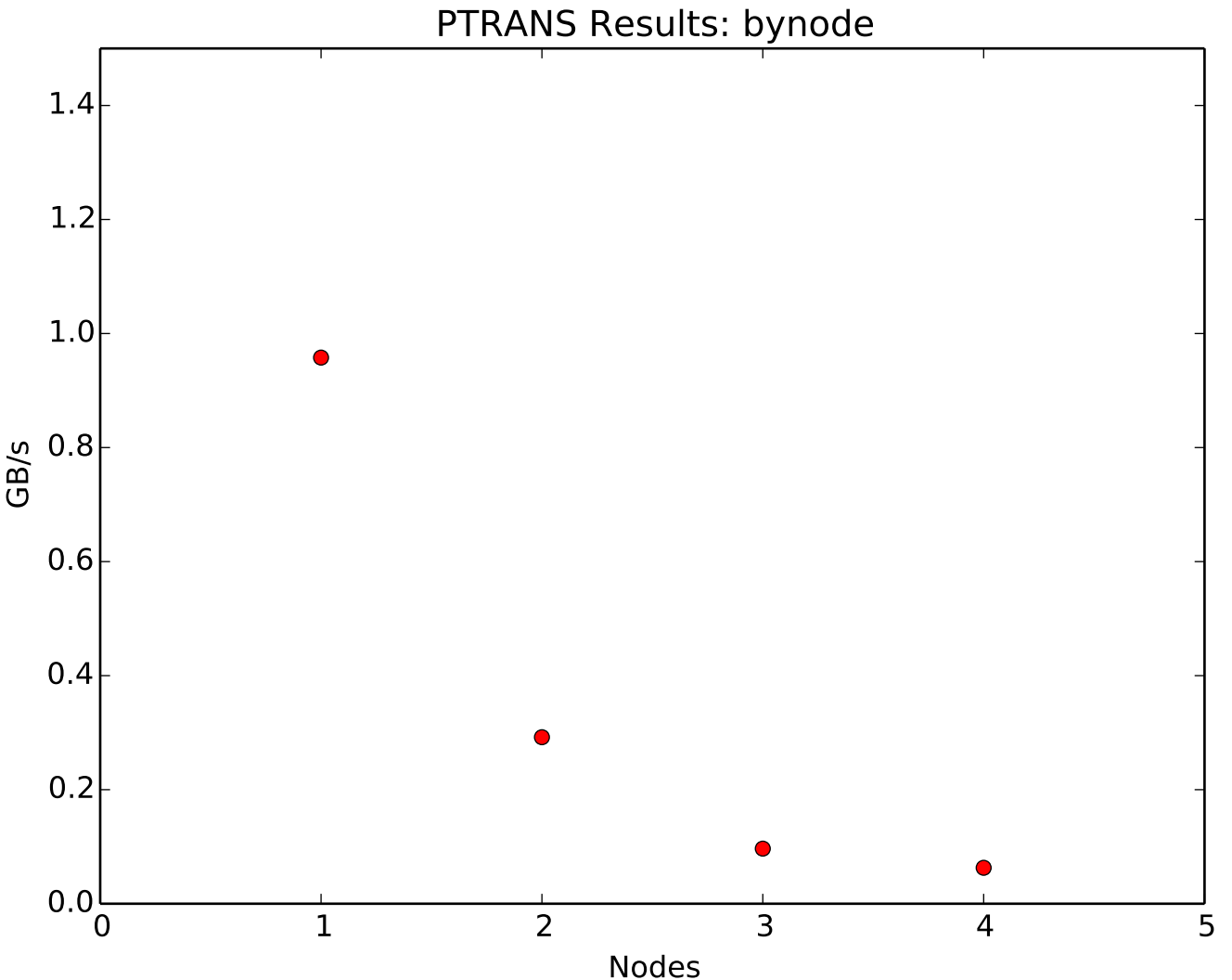

Tachyon Results: byslot
Runtime (s)
Nodes
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5

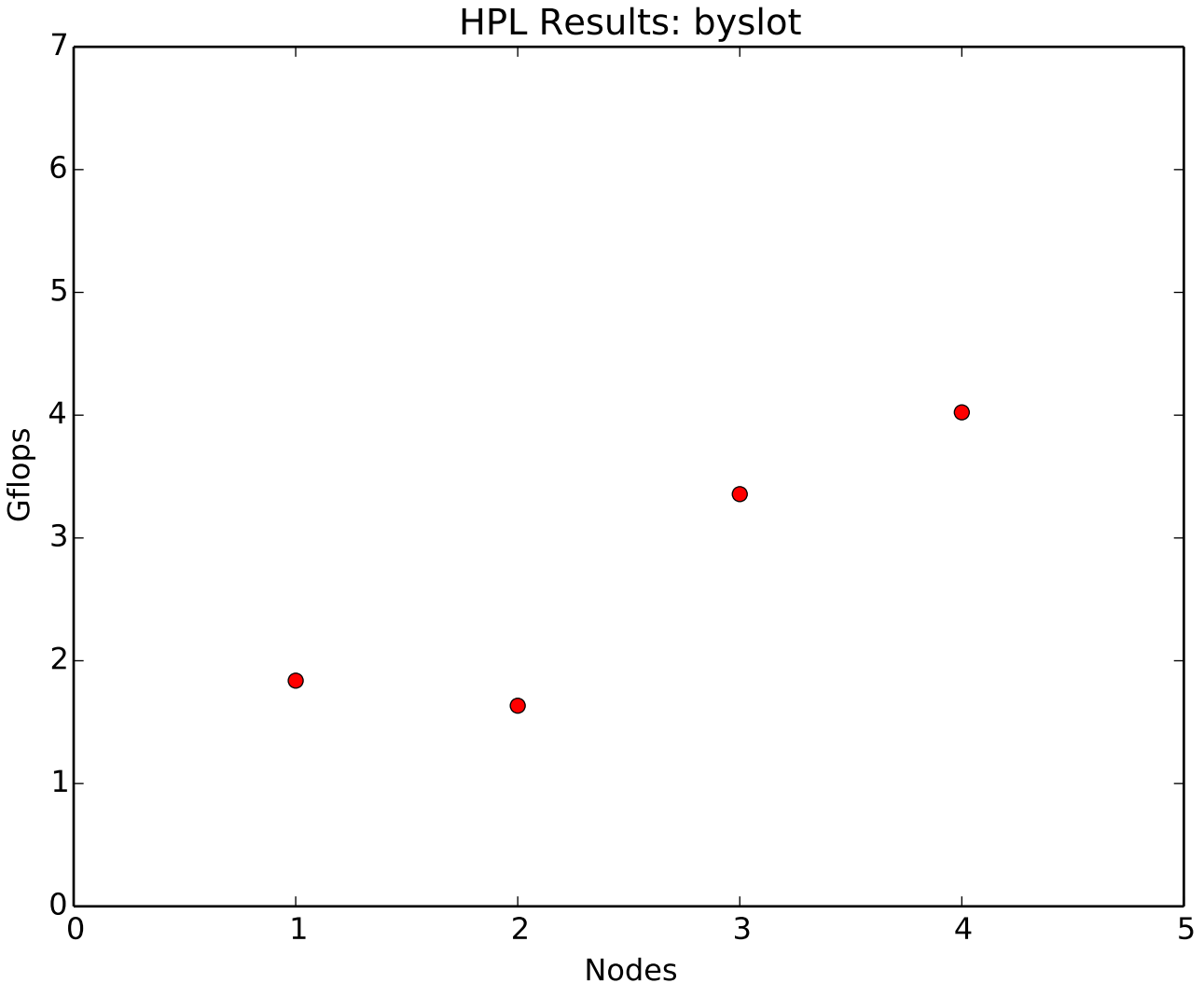

Tachyon Results: bynode
Runtime (s)
Nodes
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5

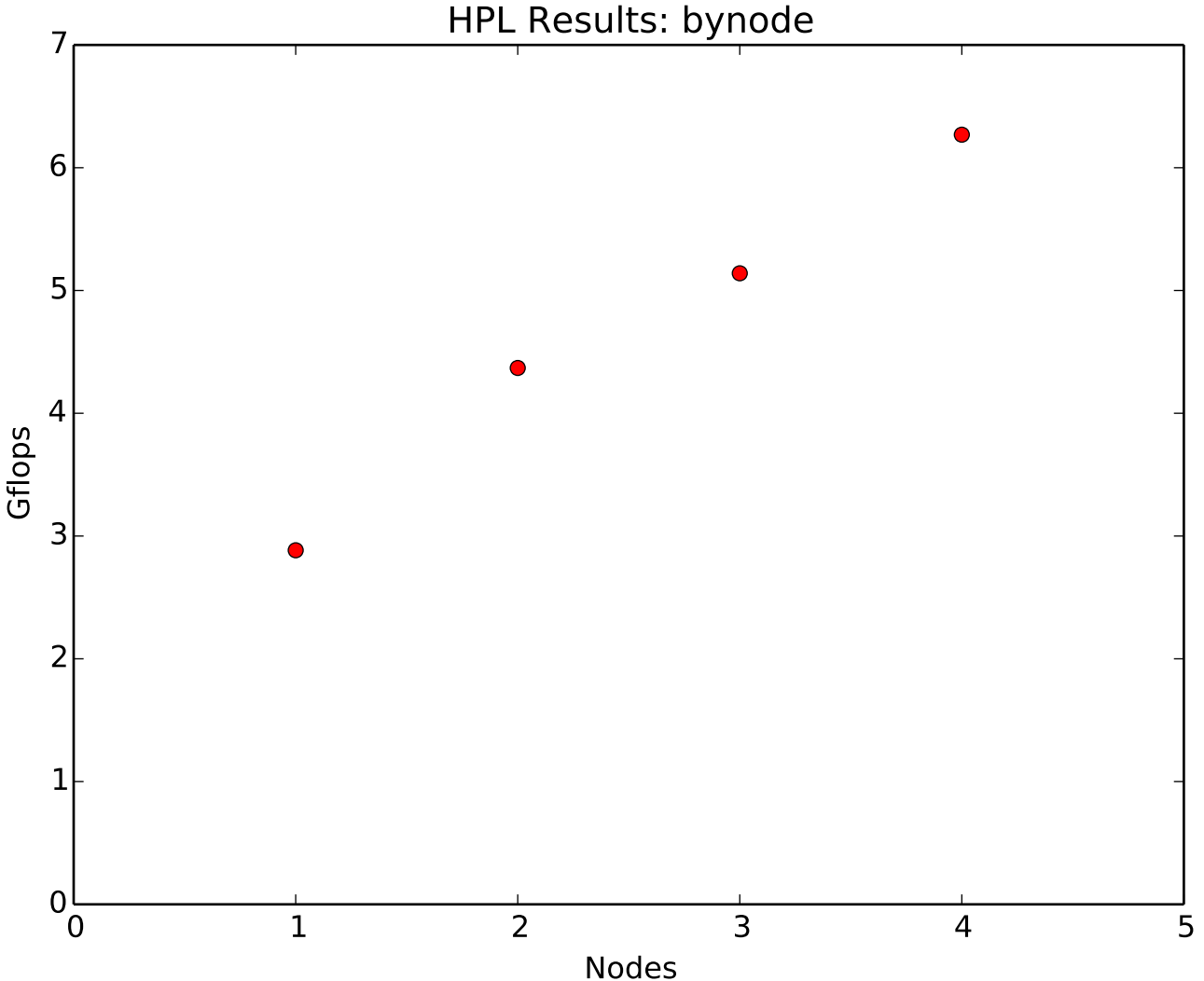
HPCC Input Parameters

| Nodes | P | Q | N | NB |
|---|---|---|---|---|
| 1 | 1 | 1 | 600 | 96 |
| 2 | 1 | 2 | 1200 | 96 |
| 3 | 1 | 3 | 1800 | 96 |
| 4 | 1 | 4 | 2400 | 96 |
| | | | | |
| N | PTRANS | HPL | VM=1 | |
| 200 | 2.2 | 1.13 | | |
| 400 | 3.3 | 1.32 | | |
| 500 | 3.7 | 1.33 | | |
| 600 | 3.7 | 1.34 | | |
| 1000 | 2.1 | 1.5 | | |
| 1200 | 0.45 | 1.82 | | |
| 1800 | 0.55 | 2.26 | | |
| | | | | |
| | | | VM=2 | |
| 800 | 0.23 | 1.12 | | |
| 1000 | 0.31 | 1.15 | | |
| 1200 | 0.38 | 1.73 | | |
| 1600 | 0.31 | 1.94 | | |
| 2000 | 0.31 | 2.26 | | |
| | | | | |
| | | | VM=3 | |
| 1200 | 0.06 | 2.57 | | |
| 1600 | 0.1 | 3.04 | | |
| 1800 | 0.08 | | | |
| 1700 | 0.1 | | | |


PTRANS Results: byslot
GB/s
Nodes
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1
2
3
4
5


PTRANS Results: bynode
GB/s
Nodes
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
1
2
3
4
5


HPL Results: byslot
Gflops
Nodes
0
1
2
3
4
5
6
7


HPL Results: bynode
Gflops
Nodes
0
1
2
3
4
5
6
7