An Analysis of NIC Resource Usage for Offloading MPI

Ron Brightwell and Keith Underwood

Scalable Computing Systems

Sandia National Laboratories

Albuquerque, New Mexico, USA

2004 Workshop on Communication Architectures for Clusters (CAC)

April 26, 2004

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Research Motivation

- Intelligent/programmable network interfaces have been shown to be beneficial for
  - Protocol offload (TOE, RTS/CTS, etc.)
  - Application offload (MPI matching, collectives, etc.)
- Typical network interface resources
  - Slower CPU relative to host CPU(s)
  - Smaller memory relative to host
- Research questions:
  - What level of resources is needed?
  - What is the impact of limited resources?

Practical Motivation

- Red Storm Seastar NIC
  - 500 MHz embedded PowerPC
  - 384 KB on-board scratch RAM
  - Other (possibly) scary things
  - Portals 3.3 programming interface
- Practical questions:
  - Will important Sandia applications work at scale?
  - What demands do Sandia applications place on network resources?
  - Will applications need to adapt to Red Storm? If so, how?

Goals

- Better understanding of how real applications use network resources
- Explore whether this type of analysis can help
  - Characterize performance/scalability of applications
  - Identify potential application performance/scalability problems
  - Determine the amount of required network resources
  - Evaluate different resource management strategies

Background

- Network resources needed by MPI
  - Processor
    - Traverse posted receive queue
      - Every time a message arrives
    - Traverse unexpected message queue (possibly)
      - Every time a receive is posted
  - Memory
    - Posted receive queue
    - Unexpected message queue (possibly)

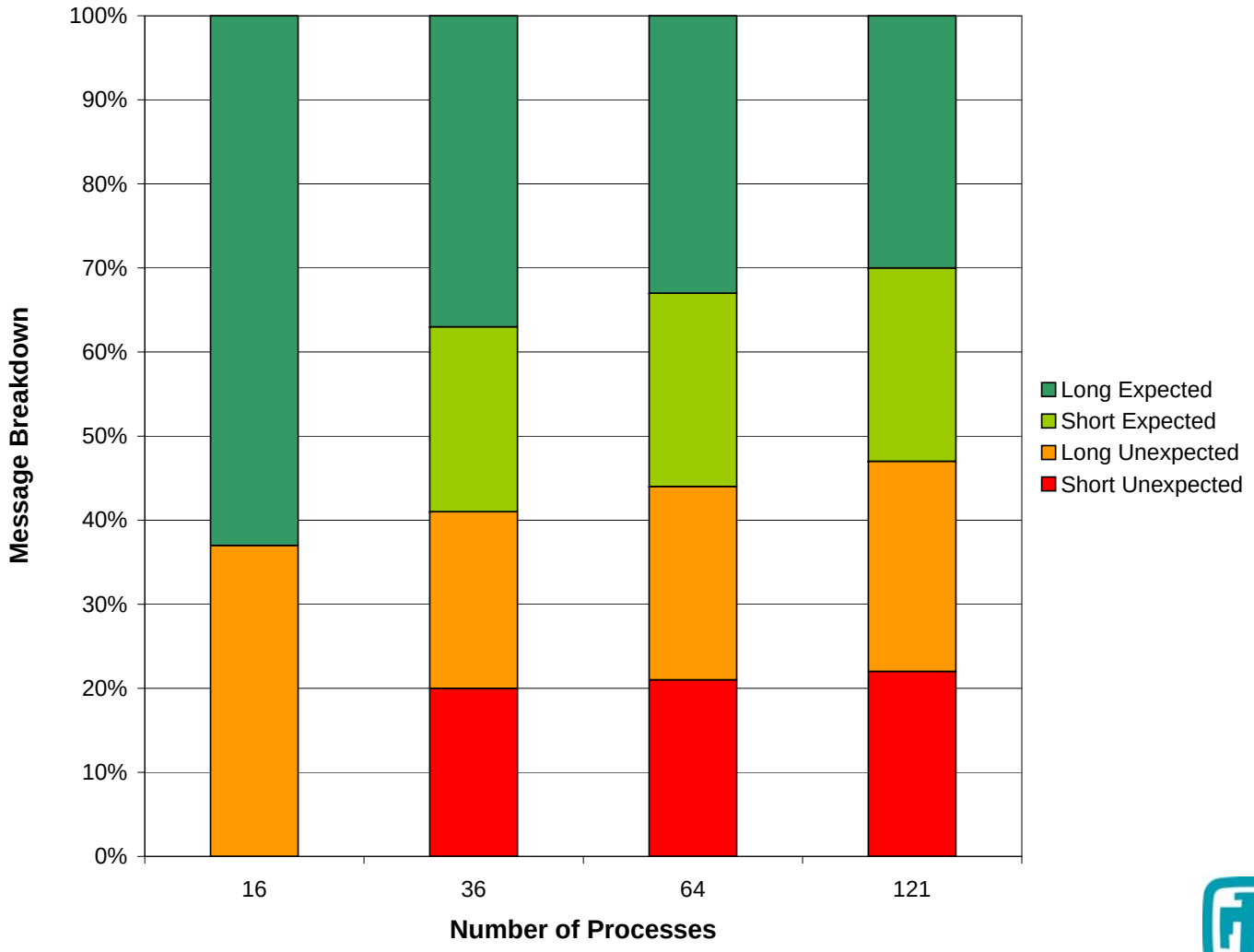
Approach

- Instrumented MPICH/GM to keep track of
  - Unexpected messages (short/long)
  - Expected messages (short/long)
  - Posted and unexpected queue data
    - Number of times searched
    - Number of entries searched
    - Maximum number of queue entries
    - Maximum number of queue entries search
- Implemented counters as global variables
- Used MPI profiling interface to write data
- Data averaged over four runs

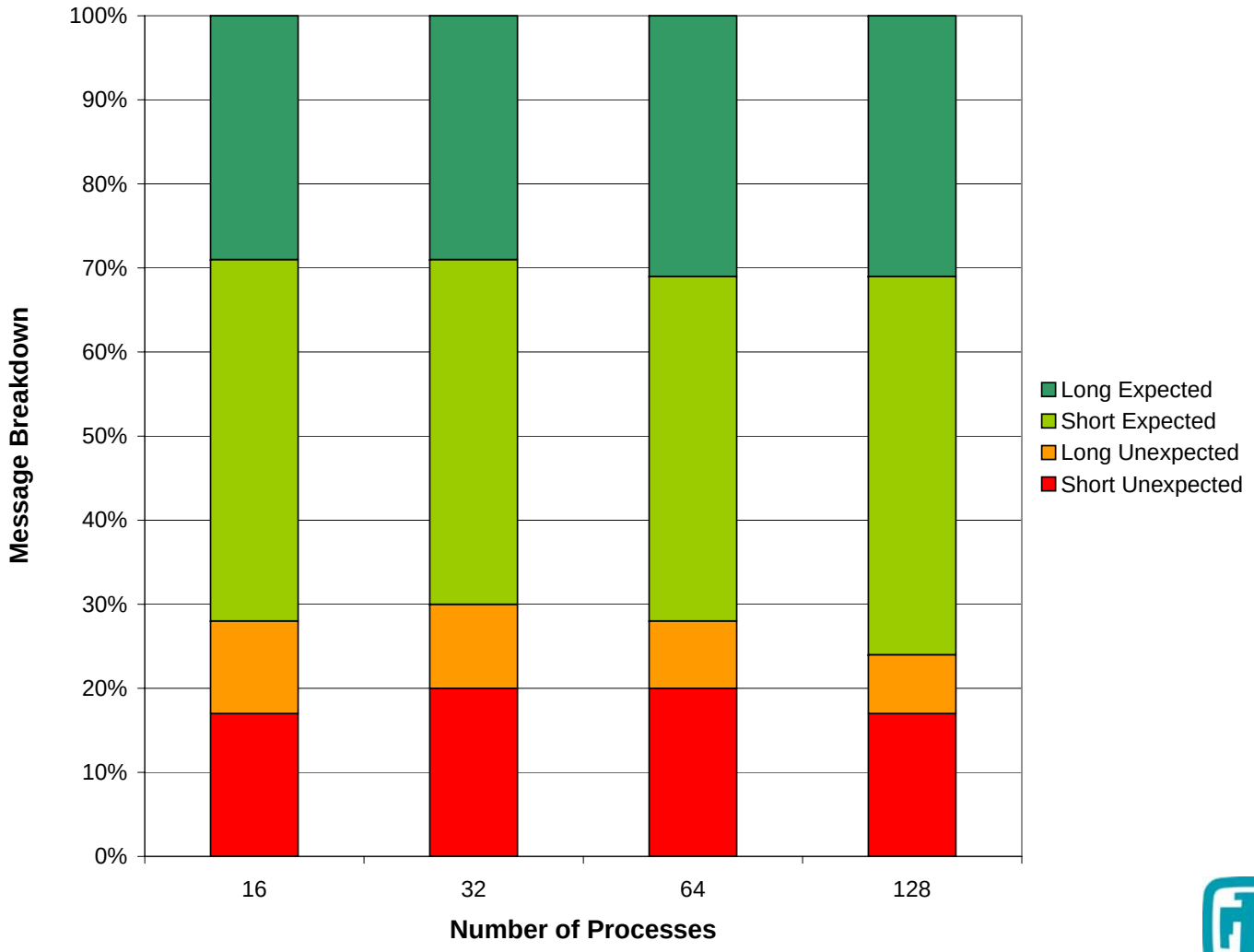
Platform and Benchmarks

- Vplant cluster at Sandia
  - Nodes
    - Dual 2 GHz Pentium IV Xeons
    - 1 GB main memory
  - Network
    - Myrinet 2000
    - Clos topology
  - Software
    - Linux 2.4.18 kernel
    - GM 1.6.4
    - MPICH/GM 1.2.4..8
- NAS parallel benchmarks 2.4 Class B (minus EP)

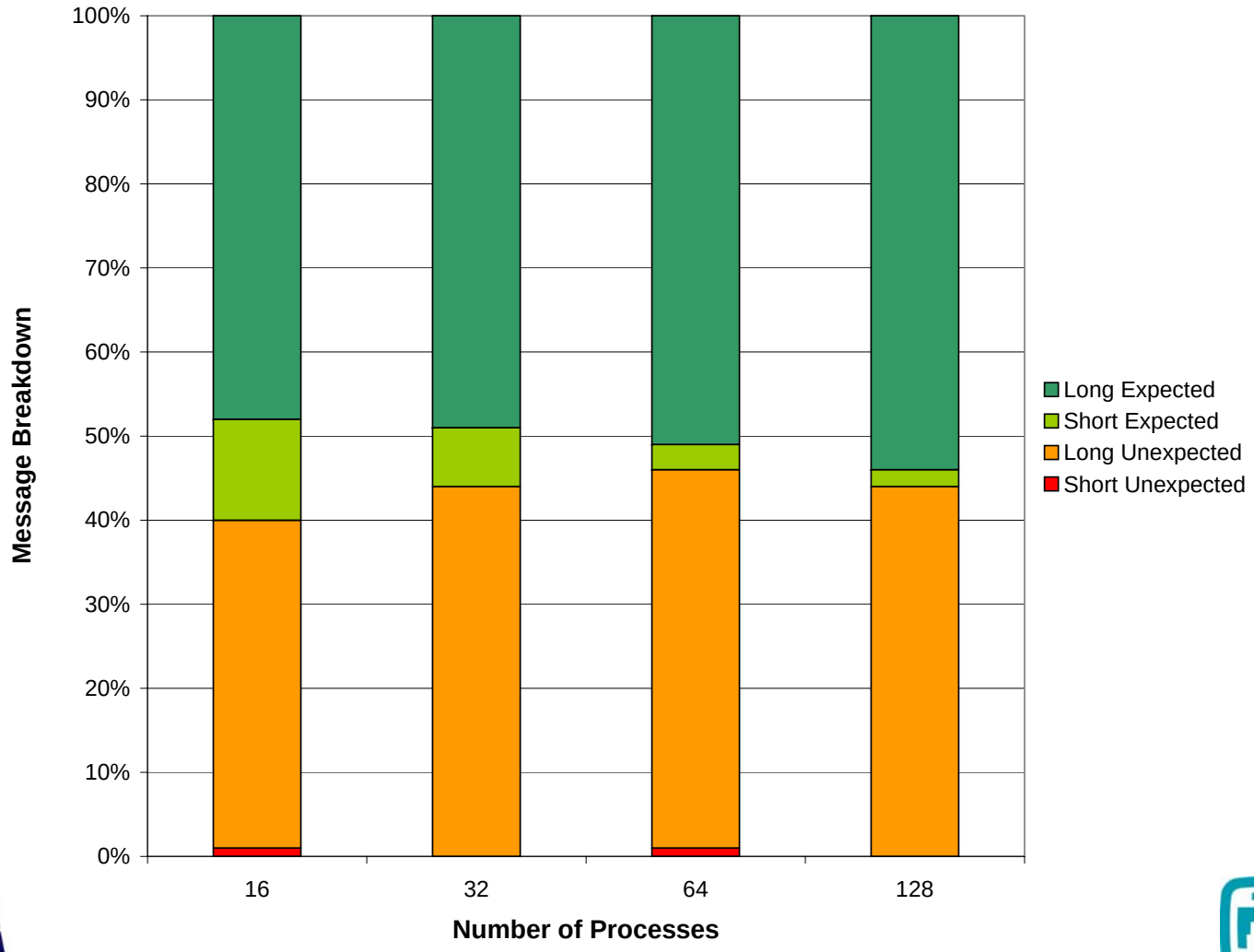
BT – Message Breakdown



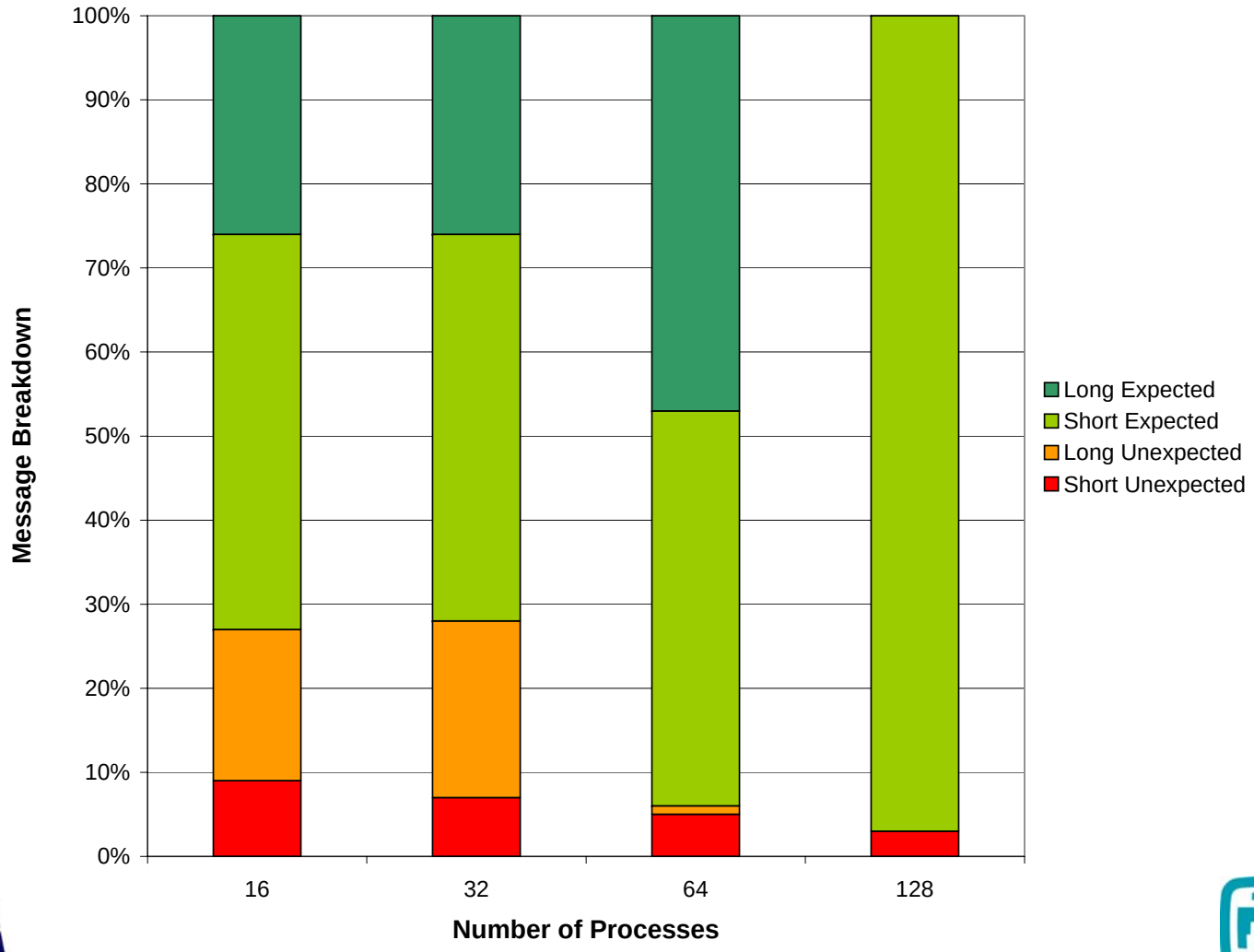
CG – Message Breakdown



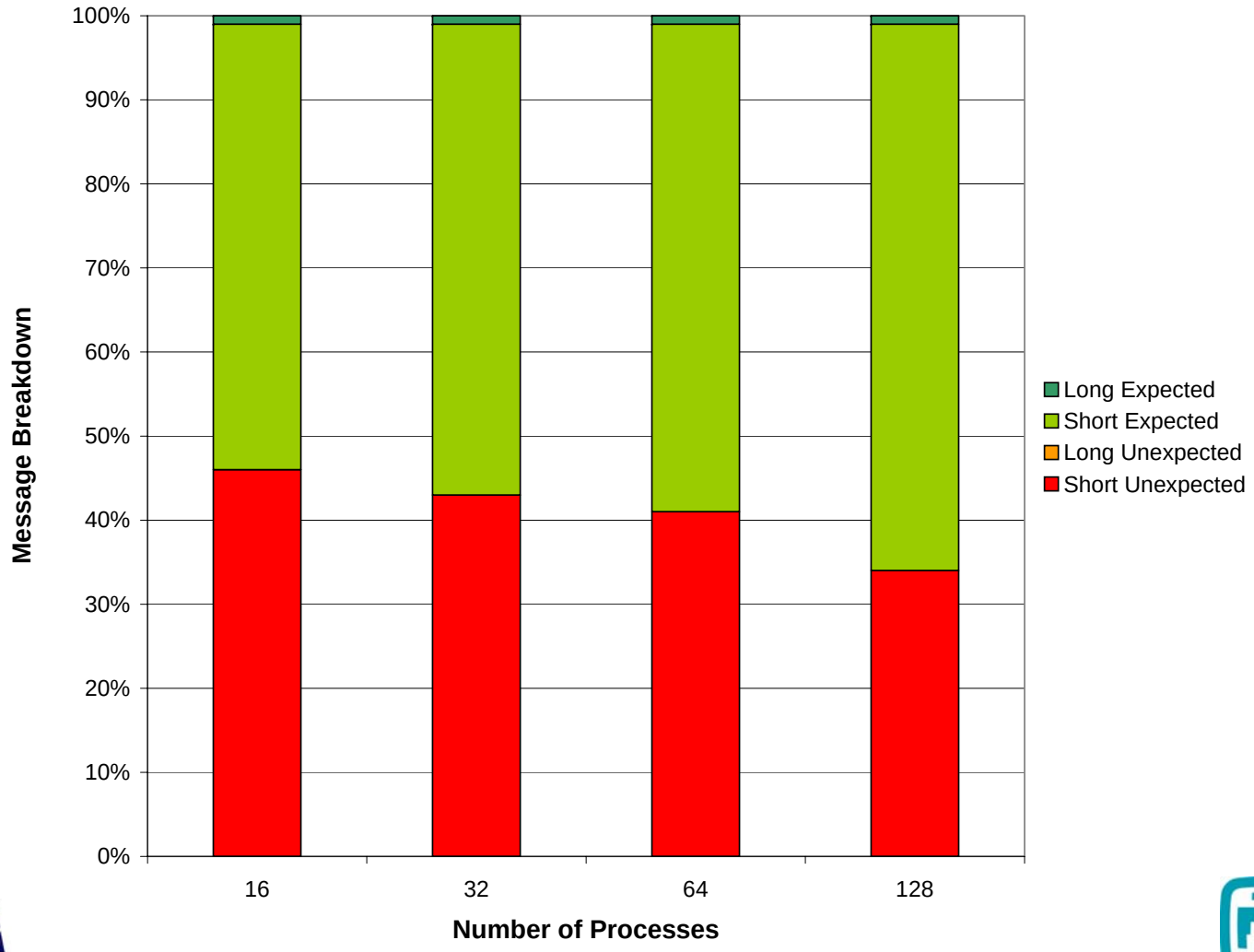
FT – Message Breakdown



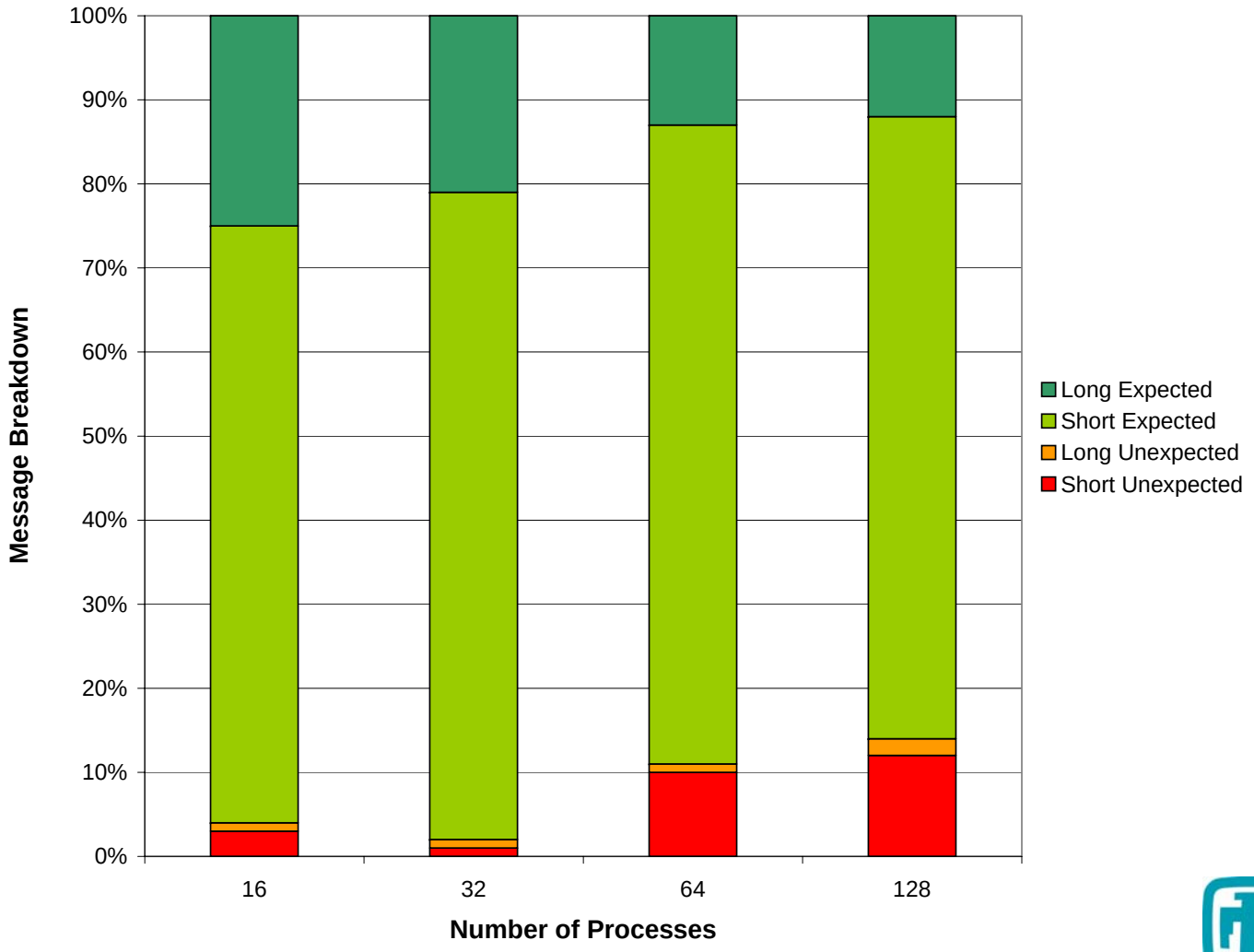
IS – Message Breakdown



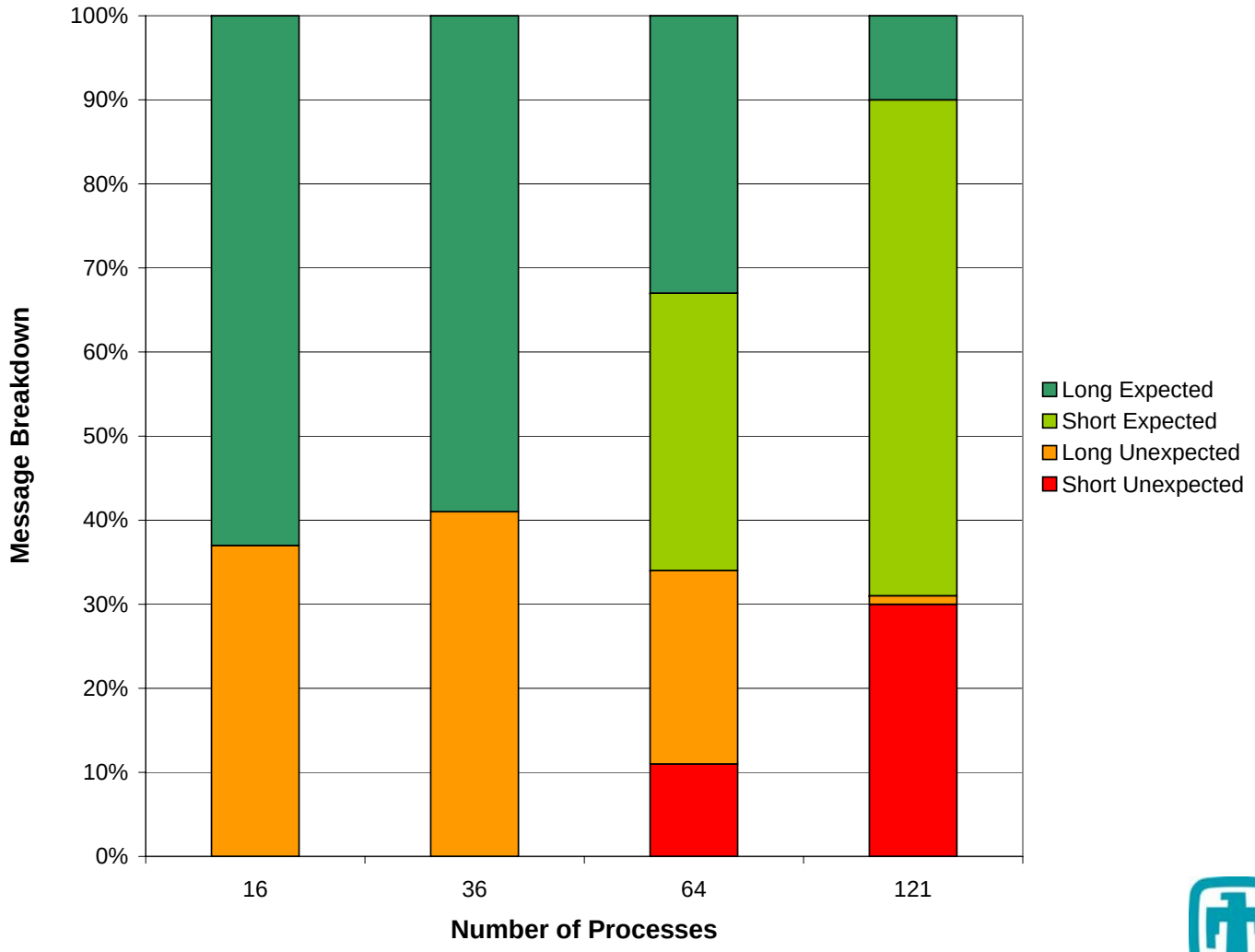
LU – Message Breakdown



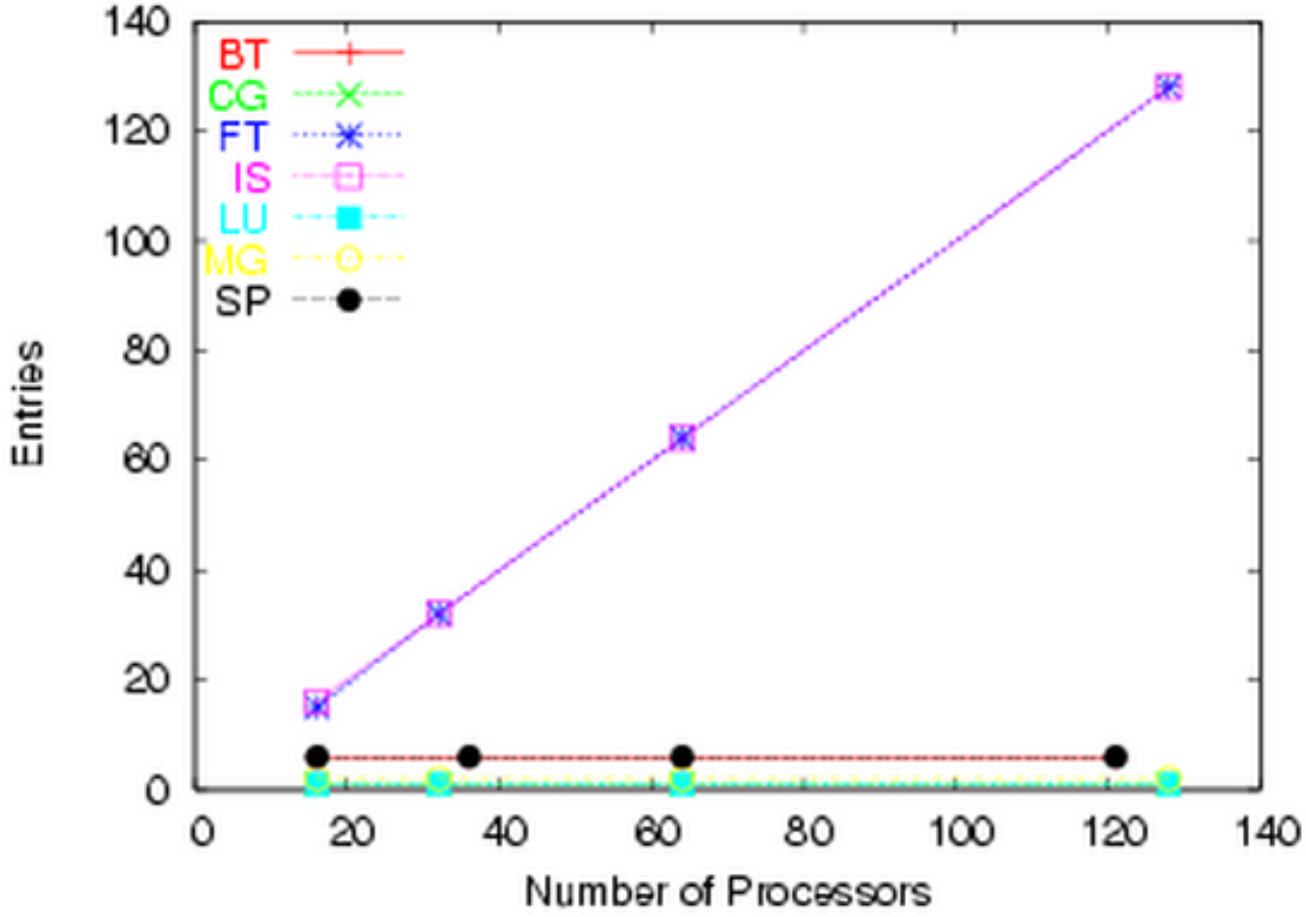
MG – Message Breakdown



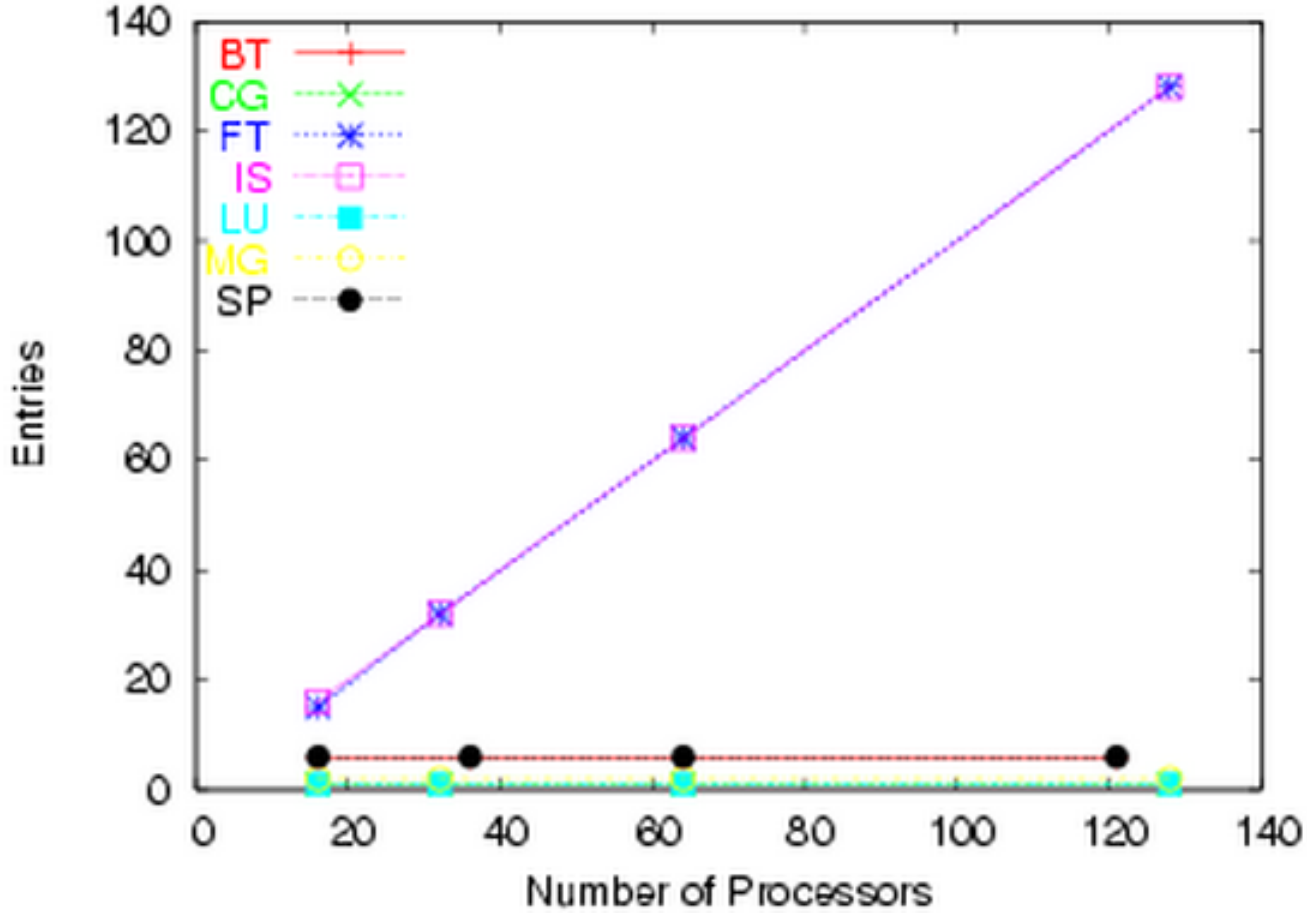
SP – Message Breakdown



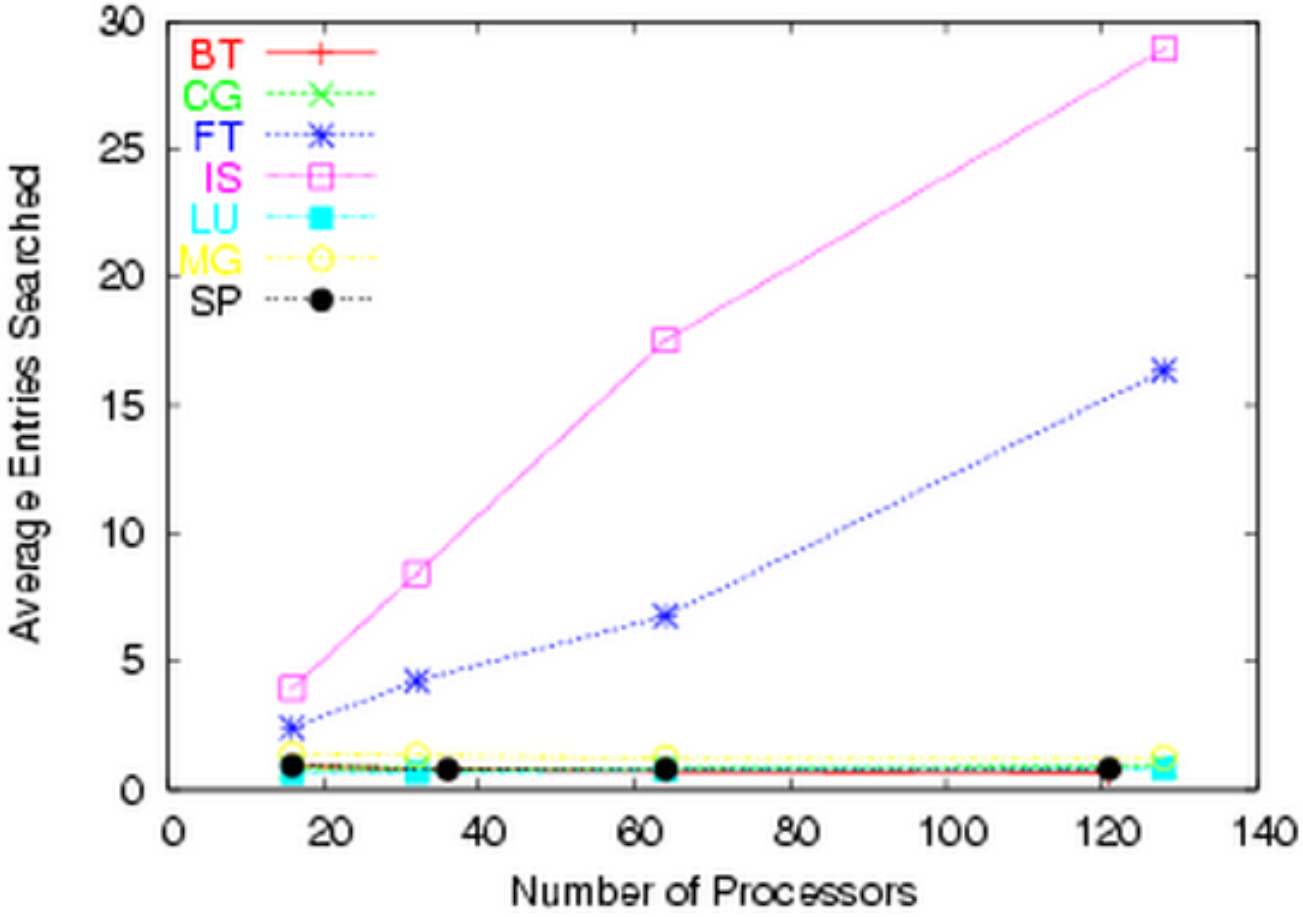
Posted Queue - Max Length



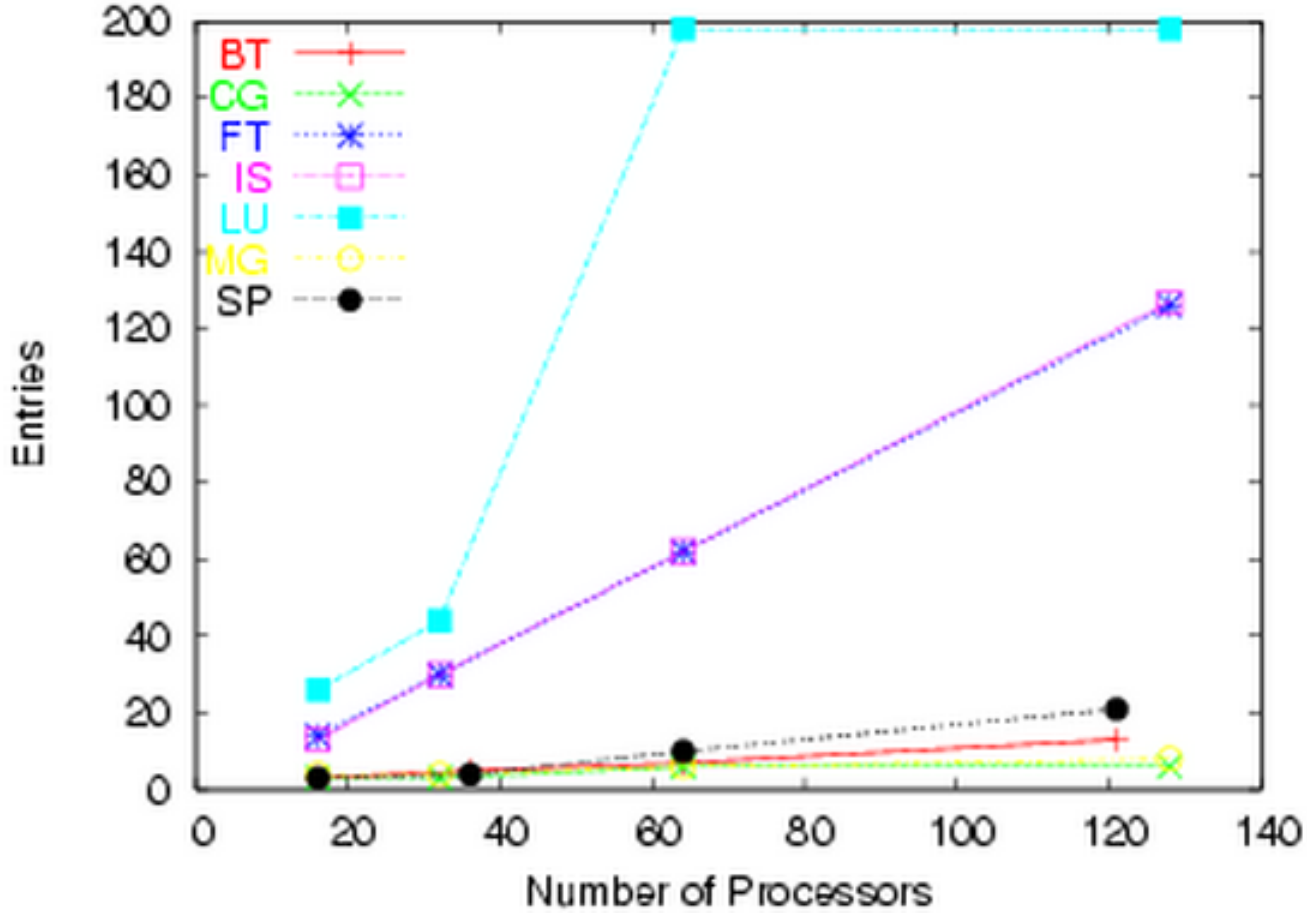
Posted Queue - Max Search Length



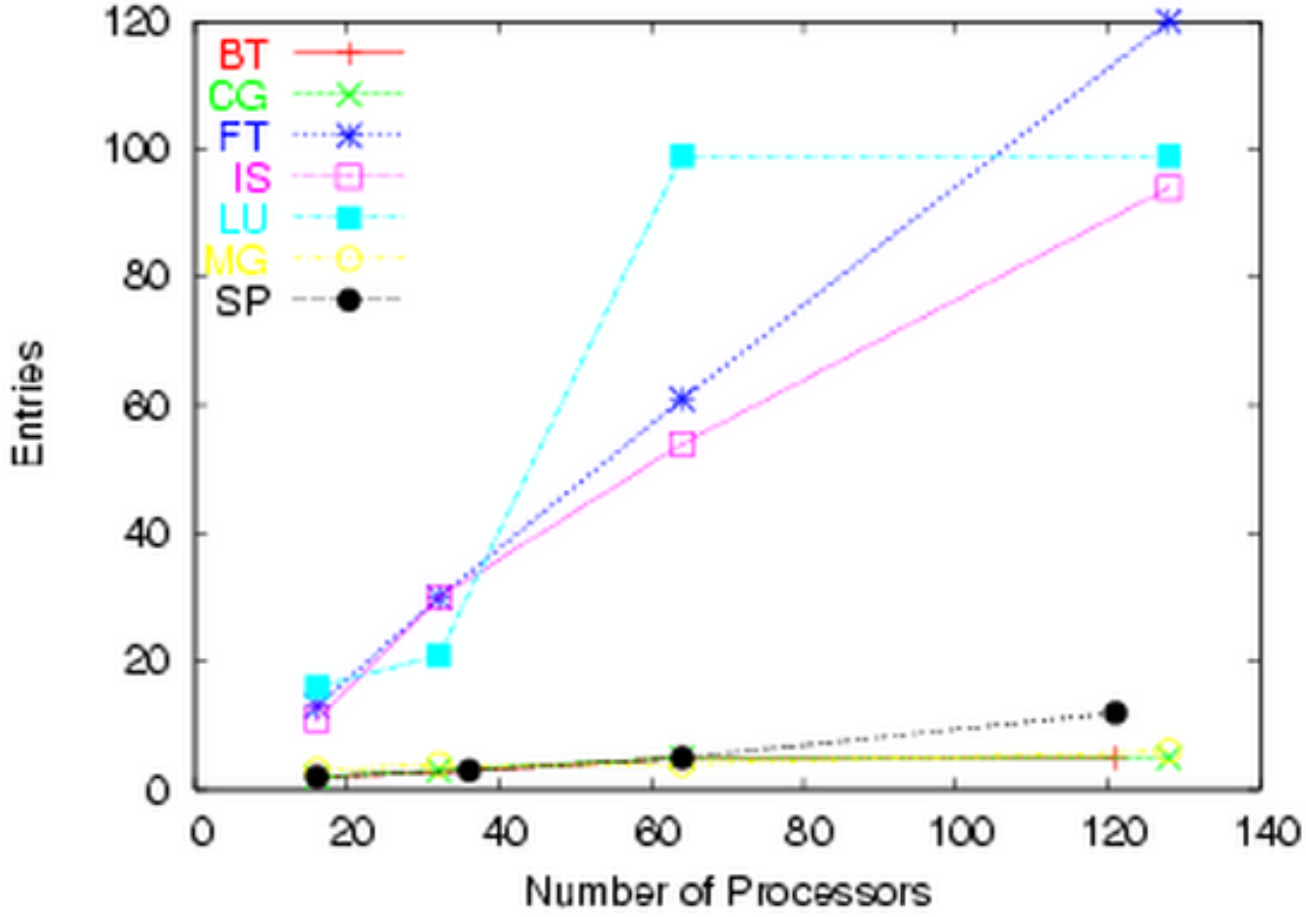
Posted Queue - Average Search Length



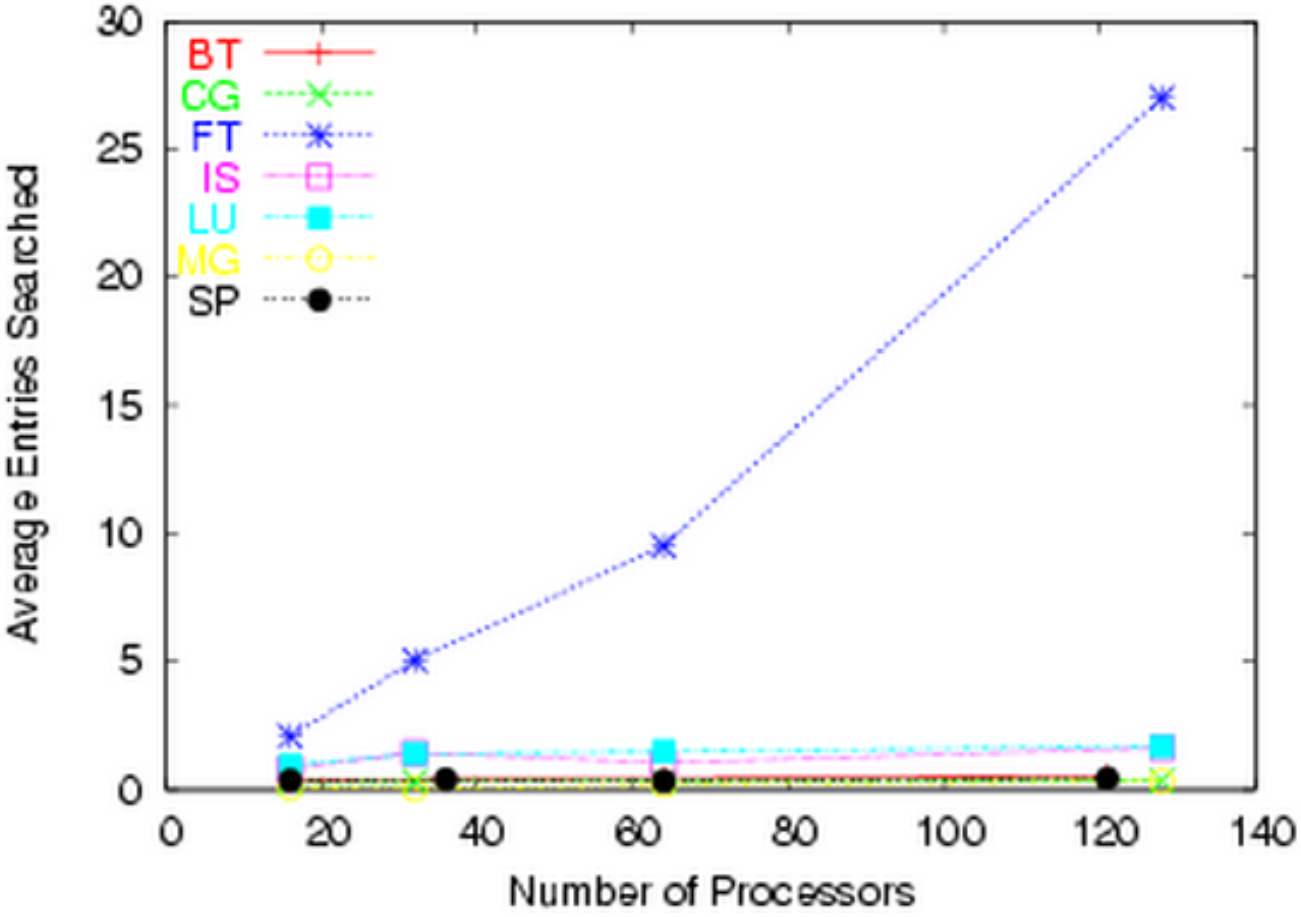
Unexpected Queue - Max Length



Unexpected Queue - Max Search Length



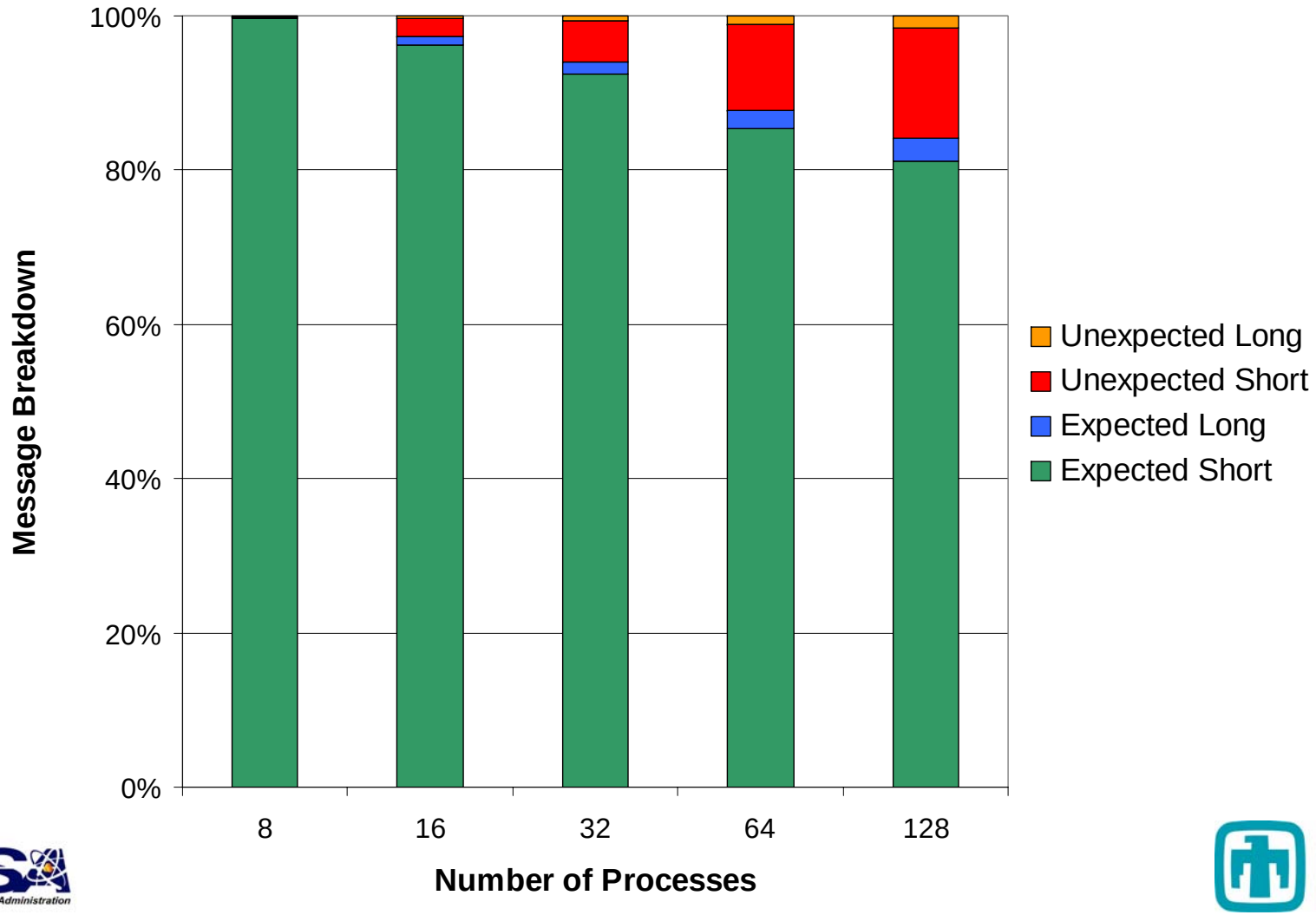
Unexpected Queue - Average Search Length



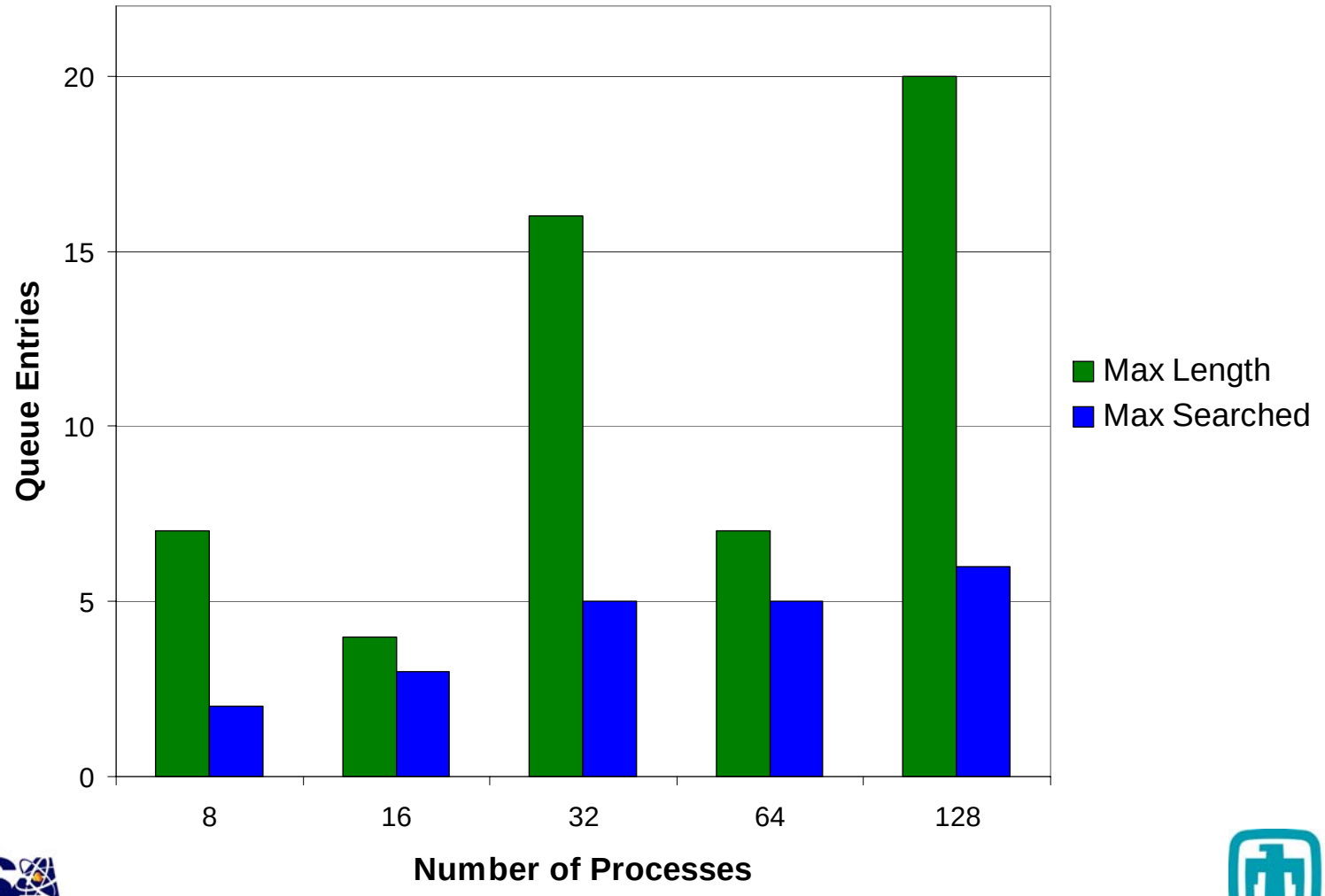
Integrated Tiger Series (ITS)

- Permits state-of-the-art Monte Carlo solution of linear time-integrated coupled electron/photon radiation transport problems with or without the presence of macroscopic electric and magnetic fields of arbitrary spatial dependence
- Results are from a typical production simulation

ITS - Message Breakdown



ITS - Unexpected Queue



Conclusions

- Usage of network resources varies dramatically across NAS parallel benchmark suite
- Significant variability in parameters for a single application
  - IS posted queue
    - Average entries searched is 30
    - Maximum entries searched is 120
- Linear growth of queues and queue traversals potentially bad
- May require greater amount of NIC resources

Future Work

- Consider more real applications
- Separate out collective statistics
- Gather data on other platforms (e.g., Elan-4)
- Design better benchmarks that test network parameters under more realistic workloads
- Understand the performance and scalability implications of resource management policies and implementations
- Better tools for data gathering and analysis