Path to Exascale Computing

Brad Benton
IBM Linux Technology Center

Date: April 15, 2010

Legal

Trademarks and disclaimers

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

IBM, the IBM logo, BladeCenter, Calibrated Vectored Cooling, ClusterProven, Cool Blue, POWER, PowerExecutive, Predictive Failure Analysis, ServerProven, System p, System Storage, System x , System z, WebSphere, DB2 and Tivoli are trademarks of IBM Corporation in the United States and/or other countries. For a list of additional IBM trademarks, please see http://ibm.com/legal/copytrade.shtml.

The following are trademarks or registered trademarks of other companies:

Java and all Java based trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries or both
Microsoft, Windows,Windows NT and the Windows logo are registered trademarks of Microsoft Corporation in the United States, other countries, or both.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries or both.
Linux is a trademark of Linus Torvalds in the United States, other countries, or both.
Cell Broadband Engine is a trademark of Sony Computer Entertainment Inc.
InfiniBand is a trademark of the InfiniBand Trade Association.

Other company, product, or service names may be trademarks or service marks of others.

NOTES:
Linux penguin image courtesy of Larry Ewing (lewing@isc.tamu.edu) and The GIMP

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Users of this document should verify the applicable data for their specific environment.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Information is provided “AS IS” without warranty of any kind.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices are suggested US list prices and are subject to change without notice. Starting price may not include a hard drive, operating system or other features. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use.

The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time.

Agenda

- **Attributes of Exascale Class Systems**
- **Exascale Class Problems**
- **Exascale Challenges**
- **Implications for OSS/Linux**

Attributes of Exascale Class Systems

Attributes of an Exascale Class system

System Peak FLOPS/OPS	**10^{18}**
System Memory	**10 PB**
Node Performance	**1-10 TF**
Storage	**300PB**
I/O	**20 TB/s**
MTBF	**1 Day**
Power	**20 MW**

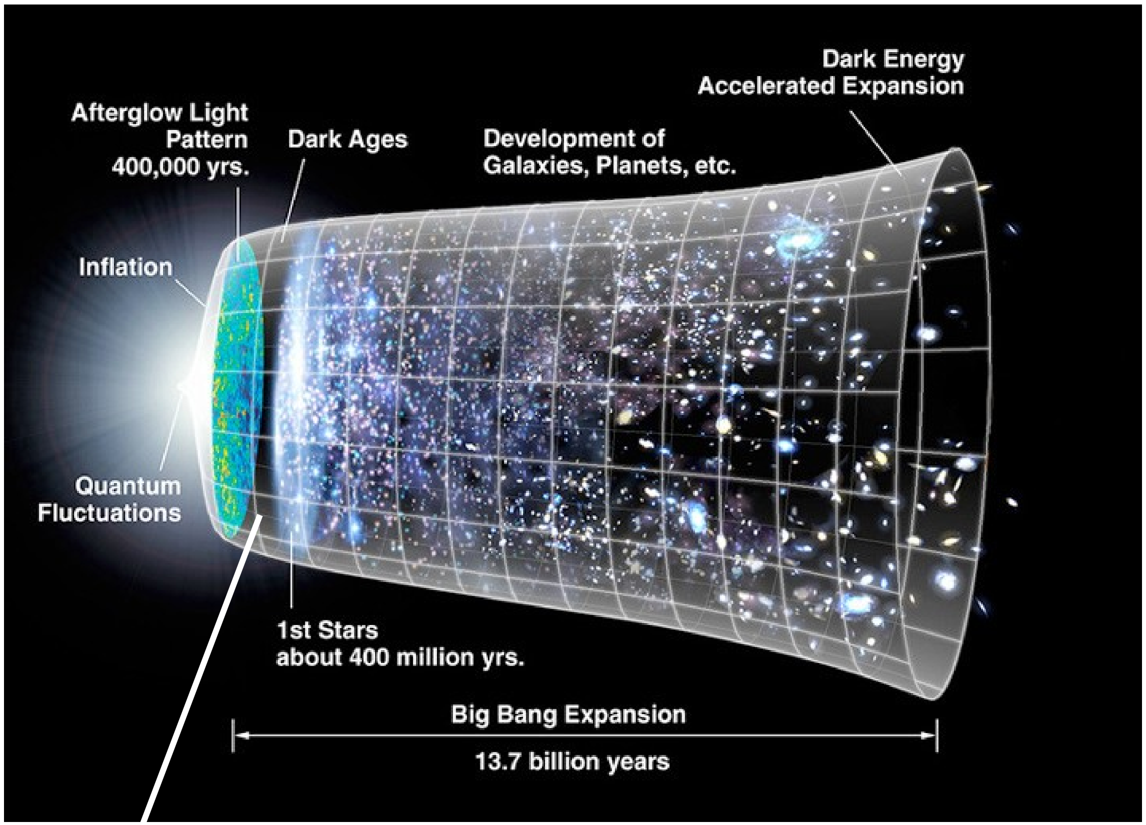
From Petascale to Exascale

If every person in the United States calculated 1 Flop/s:

- 1 PetaFLOP would take ~37 days
- 1 ExaFLOP would take ~102 years

Age of the Universe in PetaSeconds: ~432 PS

Age of the Universe in ExaSeconds: ~0.5 ES



1 ExaSecond

1 PetaSecond

~34 Billion Years

Exascale Class Problems

Exascale Problems

- **New levels of capability computing for simulations and modeling (e.g., 3D vs. 2D simulations)**

- **Increased capacity computing (e.g., multiple, simultaneous simulations to explore alternatives)**

Exascale Problems

- **Energy Research**
  - Combustion, Nuclear Fission, Solar, Nuclear Fusion…
- **Environment**
  - Climate Modeling, Multi-physics simulations
- **Biology**
  - Multiscale molecular modeling, bioinformatics, …
- **Socioeconomic Modeling**
- **Astrophysics**
  - Core-collapse supernovae, Stellar Evolution, Galaxy Formation
- **Etc., etc.**

Exascale Challenges

Moore's Law

- Can no longer rely on increasing system performance by increasing clock frequency.
- However, Moore's Law still applicable; but by doubling cores/chip every 18 months
- Cores will likely be heterogeneous: a mix of GP and Specialized

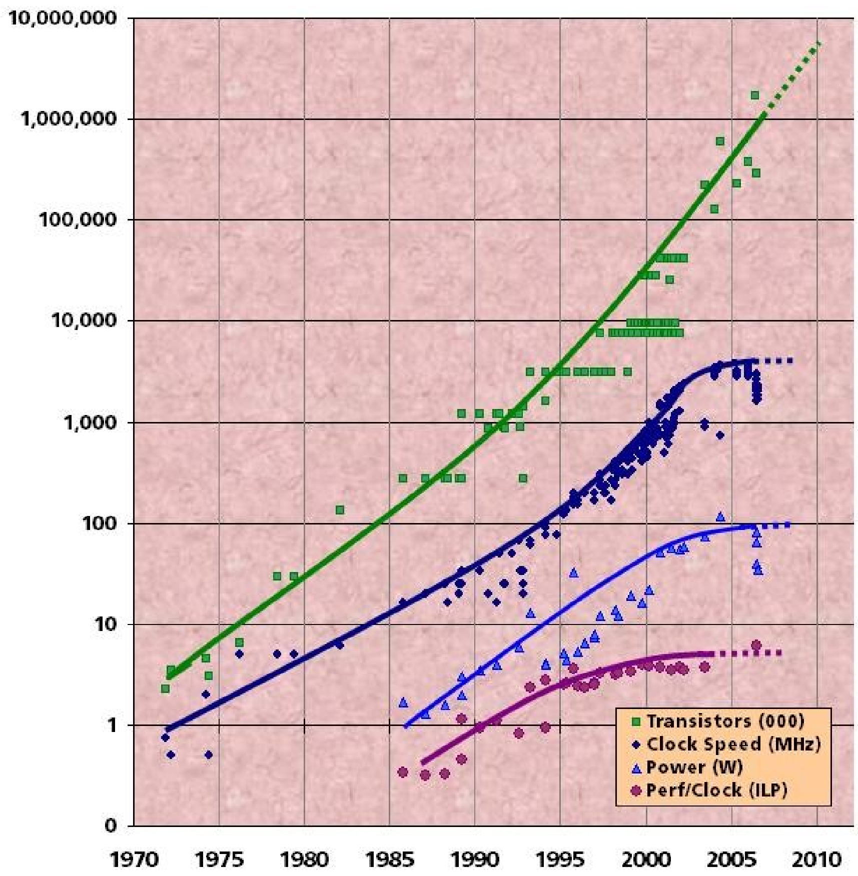


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Exascale Challenges

- **Energy and Power Challenges**
  - Extrapolation from current technology would require ~100MW (Just need a small nuclear power plant) for an Exascale system
  - Goal is for no more than 20-25 MW of sustained power consumption
- **Memory and Storage Challenge**
  - Need new Technologies
  - 3D die stacking
  - on-chip photonics
  - Phase Change Memory (PCM)
  - Memristor

Exascale Challenges

- **Concurrency and Locality Challenge**
  - Can no longer get performance gains by cranking up the clock speed
  - Path from Terascale to Petascale was relatively smooth and only needed a ~10x increase in parallelism
  - The Petascale/Exascale move will significantly increase the required level of parallelism from 10s of thousands to 100s of millions of processing elements, up to O(10^9) concurrency
  - Will require new programming models

Exascale Challenges

- **Resiliency Challenge**
  - At any given time, something in the system will be broken, in the process of breaking, or being re-integrated after repair; it will never be “whole”.
  - Principle cause of failures in HPC systems is Hardware (opposite of the situation in the commercial space).
  - Hardware will have to have some level of redundancy/recovery
  - Software will have to be able to deal with failures via integration with such technologies as CIFTS FTB

Exascale Challenges

- **Resiliency/Fault-Tolerance**
  - Software Resiliency
    - More than just checkpoint/restart
    - Containers/virtualization
    - suspend/migrate/resume
  - Example: CIFTS Fault-Tolerant Backplane
    - **C**oordinated **I**nfrastructure for **F**ault **T**olerant **S**ystems (CIFTS)
    - **F**ault **T**olerance **B**ackplane (FTB)
      - Fault aware and notification backplane for uniform event handling and notifications

Fault-Tolerant Backplane (FTB)

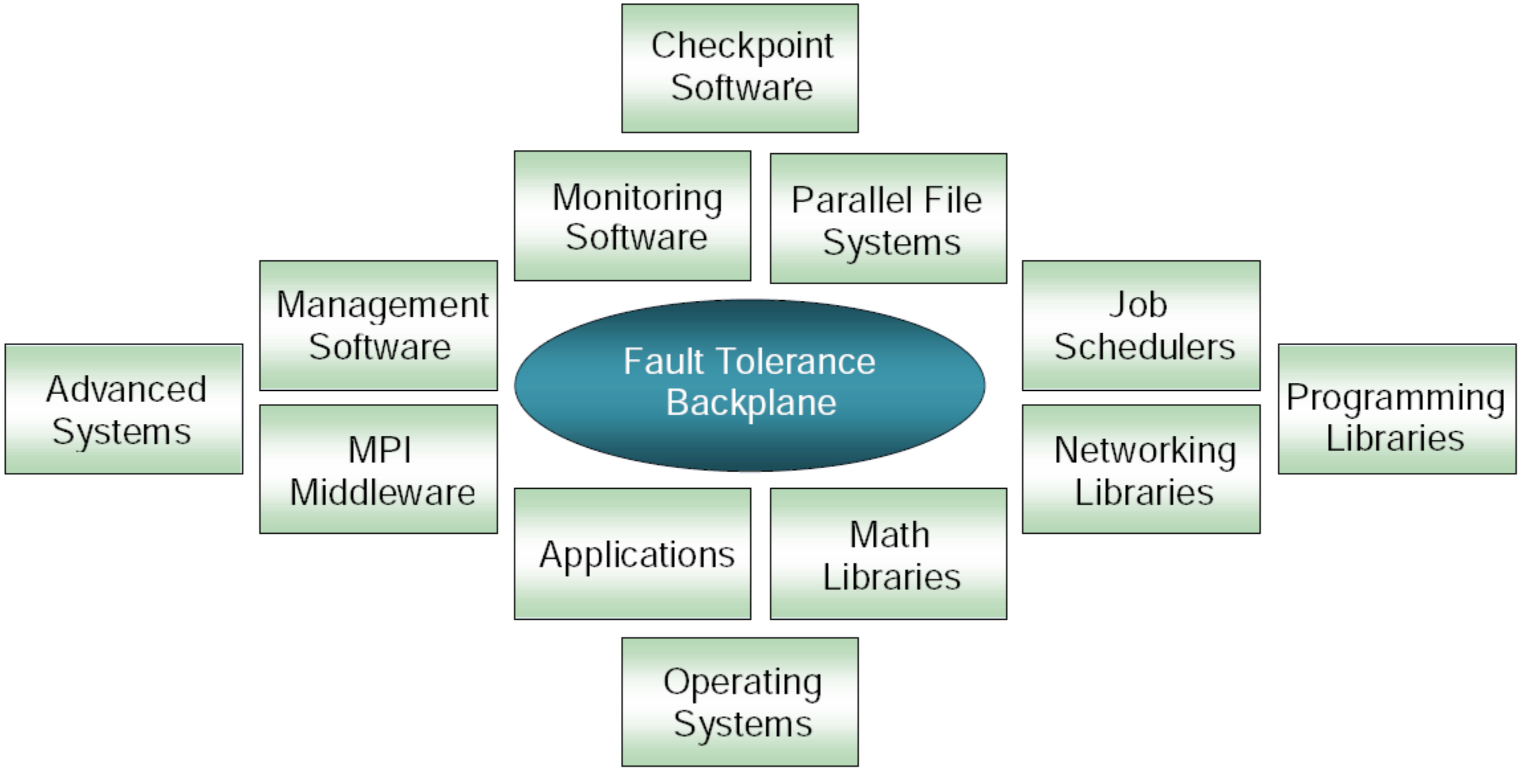


Figure courtesy of Abhishek Kulkarni

Exascale Challenges

- **Resiliency Challenge (cont.)**
  - Many-core architectures will provide for a mix of functionality, some of which can be oriented toward resilience:
    - Most cores dedicated to computational tasks
    - But other cores can be dedicated to monitoring & recovery tasks

Exascale Challenges

- **Managing 500M to 1B cores (most likely heterogeneous)**
- **Power Management**
- **Workflow Management/Process Steering**
- **Data Management/Storage/Visualization**

Programming Models for multi-core

- **MPI**
  - Will MPI survive in an exascale world?
  - 15 years of legacy code & programming experience
  - Will most likely survive in some form

- **Evolve hybrid language models: MPI +**
  - OpenMP
  - GPU Accelerators (CUDA, OpenCL)
  - PGAS languages (CAF, UPC, Chapel, Fortress, X10)
  - Need ways to coordinate resource allocation (cores/threads, affinity)
  - Models for interacting w/accelerators
  - Models for interacting w/intelligent interconnects that provide functional offload (e.g., reductions, barriers, broadcast)

Implications for OSS/Linux

Implications for OSS/Linux

- **Software will become increasingly open and dependent on a broader community**
- **Major collaborative effort across all segments: Industry, Academia, Labs**
- **HPC community has already produced an impressive list of OSS:**
  - math libraries (ATLAS, LAPACK, etc.)
  - MPI libraries
  - performance counters (PAPI, perf_events, etc.)
  - compilers, languages (Fortress, CAF, UPC, etc.)

Implications for OSS/Linux

- **However, higher level coordination of these efforts is needed to make it to Exascale**
- **The International Exascale Software Project (IESP) is attempting to provide that: www.exascale.org**

Implications for OSS/Linux: Is Linux the right OS model?

- **Some argue that it's time to move to a new, lightweight kernel for compute-specific cores**
- **However, Linux has made great strides in support of HPC**
  - Large page support
  - NUMA support
  - Read-Copy Update (RCU)
- **Selected by NCSA as the OS of choice for Blue Waters (10 PF system)**
- **And More Work is Underway**
  - OS Jitter Reduction
  - Improved management of Large pages
  - Resource Management
  - Containers (system & app)
  - perf_events
  - ummunotify (or some similar mechanism to notify userland of changes in page mappings)

Implications for OSS/Linux

- **But More is Needed**
  - Managing 100K+ processors
  - Lightweight, low-noise kernel
  - Lighter weight threads
  - Lightweight local synchronization

- **APIs for …**
  - inter/intra-node communication
  - inter/intra-node thread management
  - energy management
  - resilience

The Path to Exascale

- **Technical Evolution is not always in a straight line**
- **Different technologies evolve at different times and rates**
- **To reach exascale levels will require the consolidation and continued evolution of multiple technologies**
  - Bits a pieces of the path are already “out there”
  - Low-power embedded cores, e.g. Blue Gene
  - Specialized accelerators, e.g. use of Cell in Roadrunner, GPUs, FPGAs
  - dense packaging w/high speed interconnect, e.g. P7/IH (currently 1TF peak per single 32-core node)
  - Need to start integrating these approaches (and others) as we move forward

Post-Exascale?

Zettascale!

www.zettaflops.org

Questions?