Data analytics for IoT devices in Cloud

We developed a framework to bring data from IoT devices to a cloud environment for real-time data analysis. The framework consists of:

- Data collection nodes near the devices.
- Publish-subscribe brokers to bring data to cloud and Apache Storm coupled with other batch data processing engines for data processing in cloud.
- Our data pipe line is Robot → Gateway → Message Broker → Apache Storm.

Simultaneous Localization and Mapping (SLAM) is an example application built on top of our framework, where we exploit parallel data processing to speedup the expensive SLAM computation.

Parallel Sparse LDA

- Original LDA (orange) compared to LDA exploiting sparseness (blue).
- Note data analytics making use of Infriband (i.e. limited by communication).
- Java code running under HARP – Hadoop plus HPC pluggable.
- Corpus: 3,775,554 Wikipedia documents, Vocabulary: 1 million words; Topics: 10k topics.
- BR II is Big Red II supercomputer with Cray Gemini interconnect.
- Juliet is Haswell Cluster with Intel (switch) and Mellanox (node) infiband (not optimized).

High Performance Data Analytics with Java + MPI on Multicore HPC Clusters

We find it is challenging to achieve high performance in HPC clusters for big data problems. We approach this with Java and MPI, but improves further using Java memory maps to and off heap data structures. We achieve zero intra-node messaging, zero GC, and minimal memory footprint. We present performance results of running it on a latest Intel Haswell HPC cluster consisting 3468 cores total.