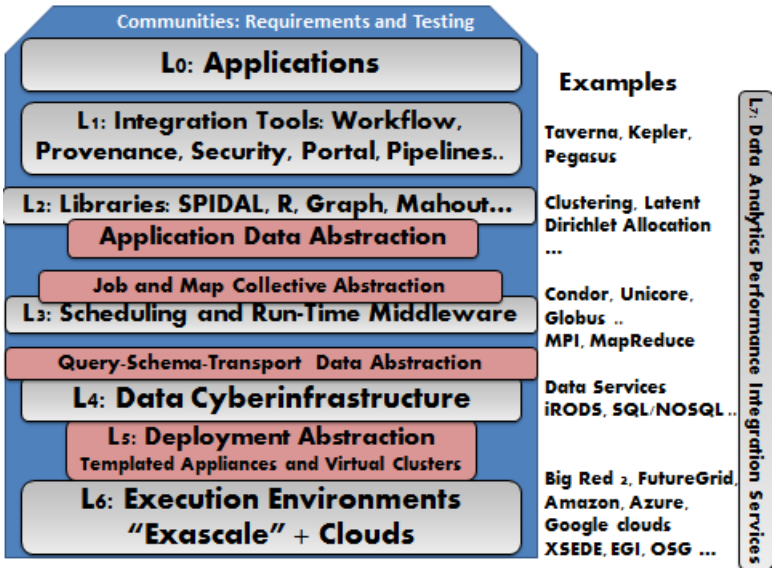
High Performance Data Analytics Ecosystem
A Proposed PTI Megaproject

Geoffrey Fox and Gregor von Laszewski



We propose the development of a High Performance Data Analytics Ecosystem aimed at data intensive scientific research and especially medical informatics. e-Commerce, Social Networking and Search applications have sprung into the forefront of data intensive applications with remarkable large systems and new (over last decade) applications and algorithms. Science has certainly participated in the data deluge but has not developed in such an impressive fashion; for example the Higgs particle was a critical discovery from the LHC but the methods used were similar to those I used as a physics experimentalist 30 years ago at Fermilab and the Grid infrastructure – although pretty big at ~200,000 cores – is not as large or broadly applicable as that used in commercial clouds. Further, I do not see much consensus as to a good research data architecture of broad applicability; much less agreement than that for the next (simulation) supercomputers. Further much of the field is using algorithms like those in R and Mahout which are not aimed at high performance whereas in simulation area improved application performance comes as much from algorithms (and libraries like PETSc, SCALAPACK, PLASMA) as from hardware; SPIDAL or Scalable Parallel Interoperable Data Analytics Library captures this requirement.

I suggest PTI collaborate on building a complete data intensive ecosystem as shown in traditional layered architecture figure where we have leading edge capabilities at all levels. It should become an internationally leading exemplar of a data intensive architecture with practical test deployments on several systems. I would suggest targeting HPC and cloud (virtualized) environments as both are likely to be important. Applications should be involved to provide requirements and test ecosystem. However they should not "lead" project.