Computational and Data Intelligence linked in the Intelligent Aether for Applications

Geoffrey Fox November 27, 2018

We are in the midst of yet another amazing computing technology-driven revolution that is seen in the advances in machine and deep learning, cloud computing, internet of things (edge computing), data and computational science. These advances are happening in spite of the slowdown in Moore's law which can be combatted by custom designs (neuromorphic or FPGA's for example), major hardware advances (e.g. quantum computing) or in the near term, by the systematic use of high-performance computing HPC technology. More importantly, we expect that the pervasive use of training and learning, the intelligent aether, can potentially lead to huge performance advances and counter the slowdown of Moore's law. All these approaches need to preserve the ease of use of current big data systems.

As well as the advances identified above, another important trend is the broadening of the importance of computing across many different application fields which is occurring within research, commercial, and government sectors. This seen academically by the broad interest in interdisciplinary programs such as "Computer Science + X", "AI Driven X", computational thinking, and of course computational and data science. Increasingly programming is moving to a higher level as users have attractive front ends (Python Notebooks or Gateways) to build mashups (or workflows) of existing libraries. There are emerging backend technologies such as Function as a Service that support this trend. The Software 2.0 concept that one programs datasets, not machine instructions, is also relevant.

We can term the systems challenge as architecting the Global AI and Modeling Supercomputer GAMSC where Global captures the need to mashup services from many different sources; AI captures the incredible progress in machine learning (ML); Modeling captures both traditional large-scale simulations and the models and digital twins needed for data interpretation; Supercomputer captures that everything is huge and needs to be done quickly and often in real time for streaming applications. The GAMSC includes an intelligent HPC cloud linked via an intelligent HPC Fog to an intelligent HPC edge. We consider this distributed environment as a set of computational and data-intensive nuggets swimming in an intelligent aether. GAMSC requires parallel computing to achieve high performance on large ML and simulation nuggets and distributed system technology to build the aether and support the distributed but connected nuggets. In the latter respect, the intelligent aether mimics a grid but it is a data grid where there are computations but typically those associated with data (often from edge devices). So unlike the distributed simulation supercomputer that was often studied in previous grids, GAMSC is a supercomputer aimed at very different data intensive AI-enriched problems.

I believe that we need to re-use as much as possible of the powerful commercial big data technology which is captured by the HPC-ABDS -- High-Performance Computing Enhanced

Apache Big Data Stack -- concept. This has been developed in a large collaborative NSF SPIDAL (Scalable Parallel Interoperable Data Analytics Library) team science collaboration project with 7 institutions led by Indiana University. This project used applications (Network Science, Polar Science, Molecular Dynamics, Pathology and Health-related Imaging) to drive HPC enhanced core technologies. Applications to scientific, medical and engineering research are attractive and these plus applications aimed at the good of society are places where academia can make especially important contributions.

There is a rapid increase in the integration of ML and simulations. ML can analyze results, guide the execution and set up initial configurations (auto-tuning). This is equally true for AI itself -- the GAMSC will use itself to optimize its execution for both analytics and simulations. This is an important idea that will grow. In principle every transfer of control (job or function invocation, a link from device to the fog/cloud) should pass through an AI wrapper that learns from each call and can decide both if call needs to be executed (maybe we have learned the answer already and need not compute it) and how to optimize the call if it really needs to be executed. The digital continuum proposed by BDEC2 is an intelligent aether learning from and informing the interconnected computational actions that are embedded in the aether. Implementing the intelligent aether embracing and extending the edge, fog, and cloud is a major research challenge where bold new ideas are needed!

The new MIDAS middleware designed in SPIDAL has been engineered to support high-performance technologies and yet preserve the key features of the Apache Big Data Software. MIDAS seems well suited to build the prototype intelligent high-performance aether. Note this will mix many relatively small nuggets with AI wrappers generating parallelism from the number of nuggets and not internally to the nugget and its wrapper. However, there will be also large global jobs requiring internal parallelism for individual large-scale machine learning or simulation tasks. Thus parallel computing and distributed systems (grids) must be linked in a deep fashion although the key parallel computing ideas needed for ML are closely related to those already developed for simulations and our expertise is immediately applicable. This technology development needs to continue being driven by applications with health, social science, and scientific research being very attractive.