Accelerating time-to-science in Fusion research

F. Spiga[1]

[1] *NVIDIA Ltd, Cambridge (UK)*

Fusion simulation codes are often the result of decades of investments in research and software developments. The majority of those codes are often capable of doing multiple kinds of complex calculations. However, not all these codes are written in a way such as efficient porting to any accelerated high-performance computing platform, like those equipped GPUs, is feasible without invasive refactoring or re-thinking entirely their structure. Current and future Exascale systems are or will be accelerated systems. Domain experts are faced with several near-existential questions: which programming model should I pick so my work will last? How can I balance specialisation and generalisation? How can I influence technology providers to adopt a useful subset of HW and/or SW features critical for my work? What concrete role AI can play in accelerate first-principle simulations?

When strictly focusing on software, the reality is there is no silver bullet the provide a clear and future-proof answer to all these questions. However there are concrete options that allow today to achieve reasonable performance improvements without sacrificing productivity and portability in software development practices. This talk aims to present NVIDIA's vision on programming models for accelerated computing, our contribution in the language and programming model ecosystem and how all of this can trigger a shift in developers' thinking from 'handling heterogeneity' to 'expose parallelism'. Few high-level examples of Fusion HPC codes will be presented aiming to demonstrate different approaches (and so different outcomes). NVIDIA is also active in pushing the boundaries of where and how ML/DL techniques can yeld a leap forward in performance and time to science, as demonstrated by few on-going research projects generating some exciting early successes.