Up to 700k GPU Cores, Kepler,
and the Exascale Future for Simulations
of Star Clusters Around Black Holes

Peter Berczik\textsuperscript{1,2,3}, Rainer Spurzem\textsuperscript{1,3,4}, Shiyan Zhong\textsuperscript{1}, Long Wang\textsuperscript{4,1},
Keigo Nitadori\textsuperscript{5}, Tsuyoshi Hamada\textsuperscript{6}, and Alexander Veles\textsuperscript{2}

\textsuperscript{1} National Astronomical Observatories of China, CAS, Beijing, P.R. China
\texttt{berczik\textsc{n}ao\textsc{.}cas\textsc{.}cn}
\textsuperscript{2} Main Astronomical Observatory, NASU, Kyiv, Ukraine
\textsuperscript{3} Astronomisches Rechen-Institut, ZAH, Univ. of Heidelberg, Germany
\textsuperscript{4} Kavli Institute for Astronomy and Astrophysics, Peking University,
Beijing, P.R. China
\textsuperscript{5} RIKEN Institute, Tokyo, Japan
\textsuperscript{6} Nagasaki Advanced Computing Center, Nagasaki, Japan

Abstract. We present benchmarks on high precision direct astrophysical N-body simulations using up to several 100k GPU cores; their soft and strong scaling behaves very well at that scale and allows further increase of the core number in the future path to Exascale computing. Our simulations use large GPU clusters both in China (Chinese Academy of Sciences) as well as in Germany (Judge/Milkyway cluster at FZ Jülich). Also we present first results on the performance gain by the new Kepler K20 GPU technology, which we have tested in two small experimental systems, and which also runs in the titan supercomputer in the United States, currently the fastest computer in the world. Our high resolution astrophysical N-body simulations are used for simulations of star clusters and galactic nuclei with central black holes. Some key issues in theoretical physics and astrophysics are addressed with them, such as galaxy formation and evolution, massive black hole formation, gravitational wave emission. The models have to cover thousands or more orbital time scales for the order of several million bodies. The total numerical effort is comparable if not higher than for the more widely known cosmological N-body simulations. Due to a complex structure in time (hierarchical blocked time steps) our codes are not considered “brute force”.

1 Introduction

Theoretical numerical modeling has become a third pillar of sciences in addition to theory and experiment (in case of astrophysics the experiment is mostly substituted by observations). Numerical modeling allows one to compare theory with experimental or observational data in unprecedented detail, and it also provides theoretical insight into physical processes at work in complex systems. Similarly, data processing of e.g. astrophysical observations comprises the use of complex
software pipelines to bring raw data into a form digestible for observational astronomers and ready for exchange and publication. Required algorithms are, for example, mathematical transformations like Fourier analyses of time series or spatial structures, complex template analyses or huge matrix-vector operations. Here fast access to and transmission of data, too, require supercomputing capacities. However, sufficient resolution of multi-scale physical processes still poses a formidable challenge, such as in the examples of few-body correlations in large astrophysical many-body systems, or in the case of turbulence in physical and astrophysical flows.

We are undergoing a new revolution on parallel processor technologies, and a change in parallel programming paradigms, which may help to advance current software towards the Exaflop/s scale and help better resolving and understanding typical multi-scale problems. The current revolution in parallel programming has been mostly catalyzed by the use of graphical processing units (GPU) for general purpose programming, but it is not clear whether this will remain the case in the future. GPU’s have become widely used nowadays to accelerate a broad range of applications, including computational physics and astrophysics, image/video processing, engineering simulations, quantum chemistry, just to name a few [10,25,24,15]. GPU’s are rapidly emerging as a powerful and cost-effective platform for high performance parallel computing. The GPU Technology Conferences held annually in San Jose (and offsprings in other parts of the world) regularly provides a snapshot of the breadth and depth of present day GPU (super)computing applications. Recent GPU’s, such as the NVIDIA Kepler K20 Computing Processor, offer 2496 CUDA processor cores and extremely fast on-chip-memory chip, as compared to only 4-8 cores on a standard Intel or AMD CPU. Groups of cores have access to very fast shared memory pieces. A single Kepler Tesla K20 device supports double precision operations fully with a peak speed of about 1 Tflop/s (double precision) and a little less than 4 Tflop/s (single precision). In this paper we use a code which still uses the single precision operations, which was originally developed for previous GPU architectures, which had no or very inefficient support for double precision. We circumvented this by emulation of a few critical parts of the code with the double precision operations “emulation” using a combination of few single precision operations (see: [22]). More details can be found in the Ph.D. thesis of one of us (Keigo Nitadori), “New approaches to high-performance N-body simulations with high-order integrator, new parallel algorithm, and efficient use of SIMD hardware”, Univ. of Tokyo, 2009.

Dynamical modeling of dense star clusters with and without massive black holes poses extraordinary physical and numerical challenges. One of them is that gravity cannot be shielded such as electromagnetic forces in plasmas, therefore long-range interactions go across the entire system and couple non-linearly with small scales. High-order integration schemes and direct force computations for large numbers of particles have to be used to properly resolve all physical processes in the system. On small scales inevitably correlations form already early

1 http://www.gputechconf.com
2 http://www.nvidia.com/gtc