

## Machines and exascale computing in LQCD

---

**Stefan Krieg,<sup>a,b,\*</sup> Eric B. Gregory,<sup>a</sup> Giovanni Pederiva,<sup>a</sup> Simon Schlepphorst<sup>a</sup> and Travis Whyte<sup>a</sup>**

<sup>a</sup>Jülich Supercomputing Centre (JSC),  
Forschungszentrum Jülich, 52425 Jülich, Germany

<sup>b</sup>Helmholtz Institute for Radiation and Nuclear Physics (HISKP),  
University of Bonn, 53115 Bonn, Germany

E-mail: [s.krieg@fz-juelich.de](mailto:s.krieg@fz-juelich.de)

We review the current landscape of high-performance computing hardware relevant to lattice QCD simulations, placing it in the context of more than four decades of co-development between the LQCD community and the HPC industry. After a brief historical survey from the quenched era through the algorithmic revolution that enabled dynamical simulations, we describe the pre-exascale and exascale systems now available to the community, with particular attention to LQCD performance results on representative machines. We then discuss JUPITER, the first exascale system in Europe, currently being deployed at the Jülich Supercomputing Centre. Looking to the future, we analyse the key hardware trends shaping the next generation of HPC: the dominance of AI workloads and its impact on vendor roadmaps, the push towards mixed- and low-precision arithmetic, the rapidly increasing energy cost of advanced semiconductor fabrication nodes, and the challenges this poses for scientific HPC users. We close with an overview of exotic and emerging hardware technologies, including RISC-V, stencil accelerators, neuromorphic computing, analog photonic computing, FPGAs, and quantum computing devices.

*The 42nd International Symposium on Lattice Field Theory (LATTICE2025)  
2–8 November 2025  
Tata Institute of Fundamental Research, Mumbai, India*

---

\*Speaker

## 1. Introduction

Lattice QCD (LQCD) simulations have always been among the most demanding computational tasks in fundamental science, and the history of the field is inseparable from the history of high-performance computing. From the very beginning, LQCD researchers both exploited and helped drive the development of the fastest available hardware. This review surveys the current HPC landscape from a lattice perspective: where we have come from, what is available today, what is being deployed at the exascale frontier, and what the coming years may bring.

## 2. HPC drives LQCD and vice versa

Since the 1980s, LQCD practitioners' hunger for compute cycles has helped to drive HPC hardware development. From the 1990s through the first decade of the 2000s this was through development of bespoke hardware dedicated to QCD calculations. Many of these machines, shown in blue in Figure 1 have been world leaders in performance.

The influence of LQCD research is also seen in the deployment of commodity HPC platforms the past two decades. In some cases QCD-specific architecture has influenced the development of commodity architecture. This is seen, for instance, in IBM's BlueGene architecture, which is widely believed to be inspired by the QCDOC [1] machine.

### 2.1 The quenched era

In the early years, only *quenched* calculations were feasible, introducing unquantifiable systematics. Even so, a complete quenched spectrum calculation required 20 years and the fastest supercomputers of the time. In 1983, a state of the art calculation was performed with  $m_N = 1000(150)$  MeV,  $m_\rho = 800(150)$  MeV,  $m_\Delta = 1300(150)$  MeV (including SU(2) and discrete approximations).

Dedicated machines built for LQCD during this period include the Columbia machines, the APE series (Italy), ACPMAPS (Fermilab), and QCDPAX (Japan). Their peak FP64 performances grew from the  $10^{-1}$  GFlops range in the early 1980s to several hundred GFlops by the early 1990s. A landmark result from this era is the GF11 [2] calculation of the quenched hadron spectrum [3], which represented roughly ten years of progress.

### 2.2 End of the quenched era and the Berlin Wall

The transition from quenched to dynamical ( $N_f = 2 + 1$ ) simulations required not only faster hardware but also algorithmic improvements. A 2002 study [4] showed that the cost of generating  $N_f = 2 + 1$  configurations diverged rapidly as the pion mass was lowered towards its physical value, giving rise to the so-called *Berlin Wall*. Hardware performance alone was insufficient; the community needed optimised integrators, multigrid solvers, Hasenbusch mass preconditioning, and multiple-integration-step-size schemes. The CP-PACS system in Japan produced a landmark unquenched result in 2000 [5]. Further dedicated machines of the era, such as QCDSF, APEmille, QCDOC, BGL<sup>1</sup>, and APEnext, brought peak performances from the TFlops to the PFlops range over roughly two decades.

<sup>1</sup>Due to their heritage, the Blue Gene machines are claimed as LQCD machines here.

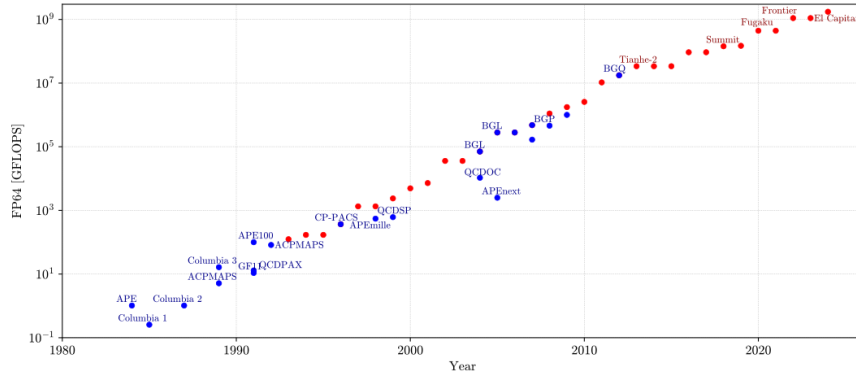


**Figure 1:** Evolution of performance leading HPC platforms over time. Red points represent the top platform from the Top 500 List [12] (since 1993). Blue points are QCD-specific machines.

### 2.3 Era of dynamical simulations and physical-point calculations

With the algorithmic breakthroughs in place and machines reaching the 100-TFlop/s scale, fully dynamical simulations became possible. Highlights include the first determination of the QCD phase-transition order at physical quark masses [6] and the first unquenched charmonium spectrum [7]. The Budapest–Marseille–Wuppertal collaboration’s calculation of the light hadron spectrum from first principles in 2008 [8] was a watershed moment, demonstrating that LQCD could reproduce the observed spectrum with controlled systematics. The first QCD+QED calculation of isospin-breaking mass splittings [9] followed within a decade.

The transition away from specialised hardware towards commodity GPU-based clusters accelerated through the 2010s, driven by the growing availability of large general-purpose HPC centres. Systems such as JUGENE (BG/P), BG/Q, and then the first GPU-based leadership facilities (e.g., Summit) replaced the community-built machines. Precision calculations such as the BMW determination of the hadronic vacuum polarisation contribution to the muon anomalous magnetic moment [10] and the development of flow-based generative samplers [11] mark the current era of precision physics and the arrival of AI-accelerated algorithms.

## 3. LQCD in the current HPC landscape

The pursuit of smaller stochastic uncertainties, smaller discretization and finite volume effects, and a denser exploration of parameter space means that LQCD researchers continue to have an immensely scalable appetite for resources. Figure 2 shows the HPC consumption for two platforms, JUWELS in Jülich and Perlmutter at NERSC, broken down by scientific domain. We can see that on these flagship platforms LQCD accounts for 20-30% of usage in one year.

This usage translates into influence on the decisions made regarding machine deployment at HPC centers, whose mission is to provide resources to its userbase. As an example, we see this explicitly in the procurement process for JUPITER, Europe’s first exascale machine. Vendor offerings were judged, in part, by performance on a suite of 23 scientific and synthetic benchmarks [15]. Among these are two LQCD benchmarks, based on Chroma [16] and dynqcd by the BMW collaboration [10].

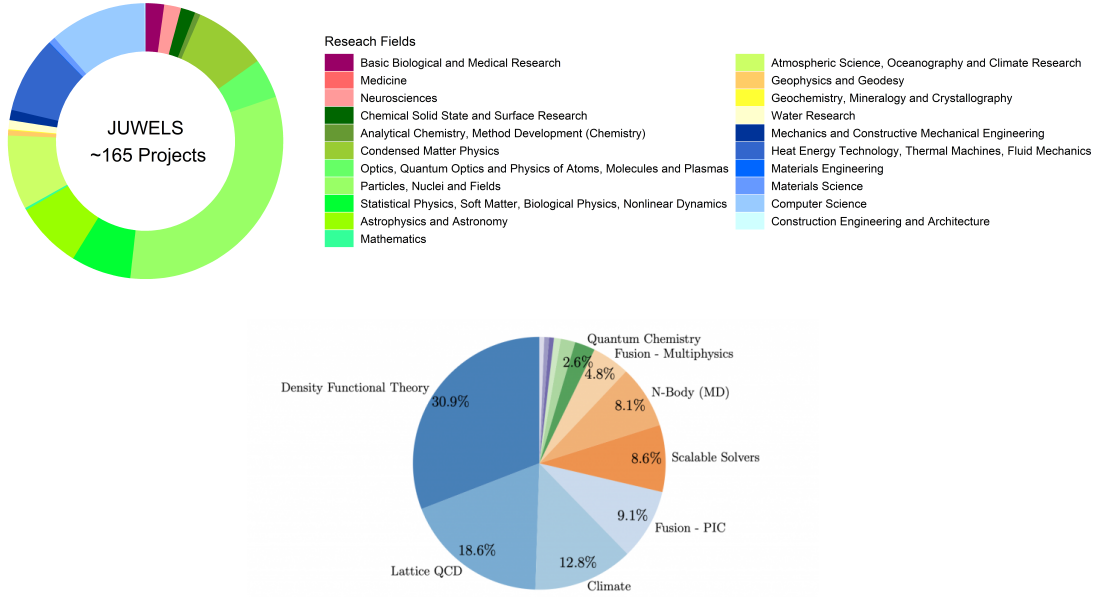


**Figure 2:** Portion of compute resources at JUWELS [13] at JSC (top) and Perlmutter [14] (bottom) consumed by different domains. At JUWELS QCD is the dominant component of the “Particles, Nuclei and Fields” contingent.

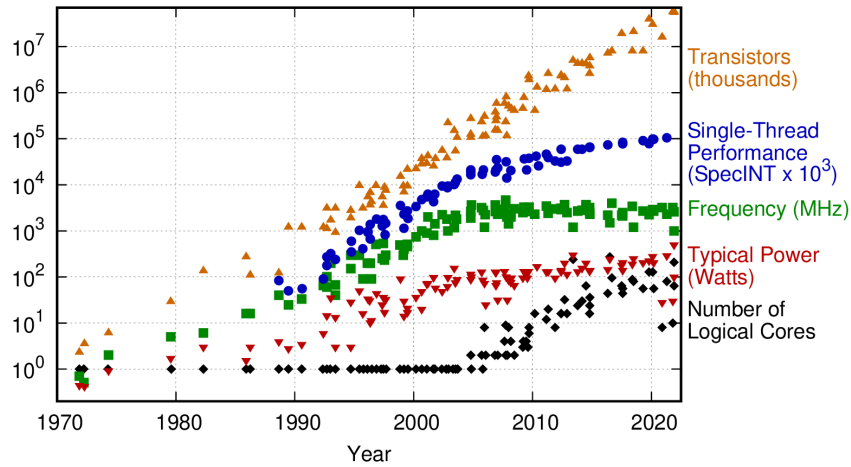


**Figure 3:** Evolution of HPC hardware metrics and deviations from Moore’s Law. [17]

### 3.1 Moore’s Law and its limits

Microprocessor performance has followed Moore’s Law for five decades. However, in the last fifteen years transistor counts continued to grow, but single-thread performance and clock frequencies have largely plateaued since around 2005–2010, while the number of logical cores has risen sharply. [17]. This shift places the burden of performance on parallelism and memory bandwidth rather than raw clock speed, with direct implications for LQCD software design.

### 3.2 The TOP500 list and accelerator landscape

As of the November 2025 TOP500 list [12], the total performance of the 500 fastest systems has reached 14.99 EFlop/s, up from 13.84 EFlop/s six months earlier, with four systems now delivering at least 1 EFlop/s each. The aggregate performance is growing at roughly one order of magnitude per decade. Of the ten fastest systems, machines with GPU accelerators dominate, with four using NVIDIA GPUs, four with AMD GPUs and one with Intel GPUs. In contrast, Fugaku [18] uses energy efficient ARM CPUs. The NVIDIA A100 and V100 hold the largest share of the accelerated performance in the list (33.9% and 17.3%, respectively, as measured in the June 2025 edition), followed by AMD MI250X and various NVIDIA H100 configurations.



**Figure 4:** Performance development of the TOP500 [12] list through the June 2025 edition (data as presented at LATTICE 2025): sum of all 500 systems (green), the #1 system (orange), and the #500 system (blue). By the November 2025 list the aggregate had reached 14.99 EFlop/s.

### 3.3 Energy efficiency

A comparison of energy efficiency (GFlops/Watt on HPL) across the current TOP500 (Fig. 5) shows that modern GPU-based systems (NVIDIA H100, GH200, AMD MI300A) occupy the upper-right corner of the performance–efficiency plane, achieving  $\sim 55$ – $64$  GFlops/Watt. The Japanese supercomputer Fugaku (ARM A64FX, #7 in the current list) achieves excellent HPCG efficiency (#2 globally) and good FP64 HPL efficiency ( $\sim 15$  GFlops/Watt), illustrating that alternative architectures can be competitive. Embedded and desktop ARM systems (Apple M4 Max, Raspberry Pi 5) occupy a middle tier.

The AI ecosystem’s focus on training-time efficiency has driven GPU vendors to invest massively in low-precision tensor-core hardware, at the expense of FP64 capability. The table at left

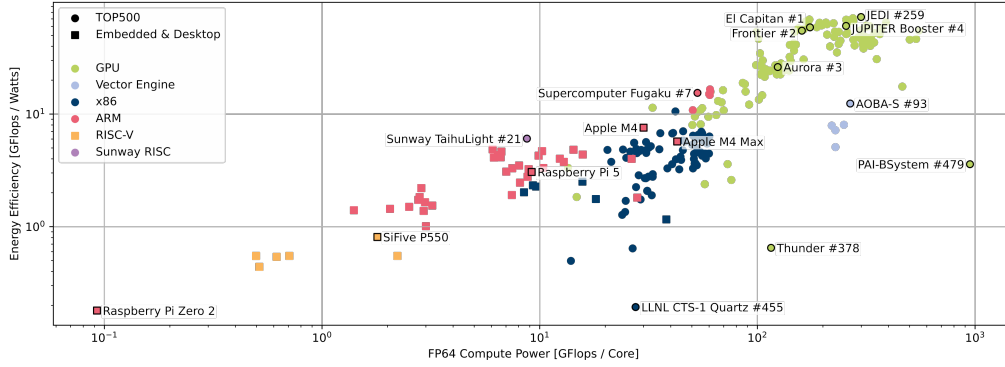


**Figure 5:** Energy efficiency (HPL GFlops/Watt) vs. FP64 compute power per core for TOP500 systems (circles, list of June 2025 [12]) and embedded/desktop devices (squares).

| Precision | Operation       | Energy/FLOP |
|-----------|-----------------|-------------|
| FP64      | FMA             | 2.5x        |
| FP32      | FMA             | 1.0x        |
| FP16      | FMA             | 0.5x        |
| FP64      | Tensor Core MMA | 1.5x        |
| FP16      | Tensor Core MMA | 0.12x       |
| FP8       | Tensor Core MMA | 0.06x       |
| INT8      | Tensor Core MMA | 0.04x       |

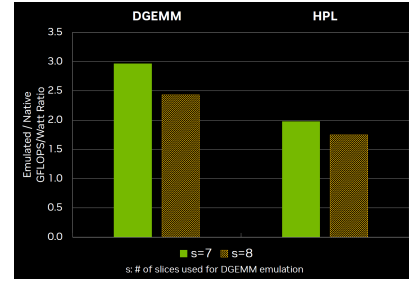


**Figure 6:** A comparison of energy consumption per flop in matrix-multiply operations between different precisions on recent NVIDIA GPUs [19].

in Figure 6 shows that recent Nvidia hardware (Blackwell) delivers the most efficient performance at low precision, with INT8 and FP8 MMA benchmarks nearly 40 times more energy efficient per FLOP as FP64. However the extra energy cost of native double precision arithmetic can be mitigated by using floating point emulation.

#### 4. Pre-exascale systems

**Fugaku (ARM A64FX) – RIKEN, Japan** Fugaku [18] comprises 158,976 nodes each hosting a single Fujitsu A64FX processor with 48 ARM cores and 32 GiB HBM2 at 1024 GB/s, interconnected via Tofu D (28 Gbps  $\times$  2 lanes  $\times$  10 ports). The system delivers 442 PFlop/s on HPL (#7) and 16 PFlop/s on HPCG (#2), on the November 2025 lists. The A64FX architecture offers outstanding memory bandwidth, making it well suited for bandwidth-limited stencil operations. Scaling benchmarks of Wilson clover, staggered, and domain wall operators with the Bridge++2.0 library [20] show near-linear strong scaling from 16 to 500 nodes.

**LUMI-G (AMD MI250X) – CSC, Finland** LUMI-G [21] has 2,978 nodes each with four AMD MI250X GPUs (8 GCDs per node), a 64-core AMD EPYC 7A53 “Trento” CPU, 512 GiB DDR4 memory, and HPE Slingshot-11 / Infinity Fabric interconnect. It delivers 379.7 PFlop/s on HPL (#9) and 4.6 PFlop/s on HPCG (#5), on the November 2025 lists. The Extended Twisted Mass

Collaboration (ETMC) has demonstrated strong scaling of tmLQCD+QUDA on LUMI-G for lattices of size  $112^3 \times 224$  [22], reaching factors of roughly  $2\times$  speedup from 28 to 112 nodes.

**Leonardo (NVIDIA A100) – CINECA, Italy** The Leonardo Booster [23] module at CINECA (Italy) contains 3,456 nodes with  $4\times$  NVIDIA A100 custom GPUs (124 SMs), one Intel Ice Lake Xeon Platinum 8358, 512 GiB DDR4, and quad-rail HDR100 InfiniBand. Its peak performance is 241.2 PFlop/s (HPL #10) and 3.11 PFlop/s (HPCG #10), on the November 2025 lists. A benchmarking study of the MILC code on this system [24] showed a time-to-solution of 178 s and an energy-to-solution of 0.56 kWh on 12 nodes.

**Perlmutter (NVIDIA A100) – NERSC, USA** Perlmutter [14] at NERSC has 1,792 GPU nodes with  $4\times$  A100 GPUs (40 GB, with 256 nodes at 80 GB), an AMD EPYC 7763 CPU, 256 GiB DDR4, and  $4\times$  HPE Slingshot-11 interconnect. It achieves 79.2 PFlop/s (HPL) and 1.9 PFlop/s (HPCG); it ranked #30 and #11 on the respective November 2025 lists. The HotQCD collaboration demonstrated near-ideal scaling of the HISQ Dirac operator in SIMULATEQCD on Perlmutter from 2 to 128 GPUs [25].

**JUWELS Booster (NVIDIA A100) – JSC, Germany** JUWELS Booster [26] at JSC has 936 nodes each with  $4\times$  NVIDIA A100 40 GB GPUs paired with AMD EPYC 7402 CPUs (2 sockets, 24 cores each), 512 GiB DDR4 and  $4\times$  Mellanox HDR200 InfiniBand ConnectX 6. It delivers 44.1 PFlop/s (HPL) and 1.3 PFlop/s (HPCG); it ranked #48 and #14 on the respective June 2025 lists. PUNCH4NFDI scaling tests of domain-wall fermions with the Grid library [27] show good strong and weak scaling to 32 GPUs.

## 5. Exascale systems

**El Capitan (AMD MI300A) – LLNL, USA** El Capitan [28] at Lawrence Livermore National Laboratory is the current world number-one system, using AMD Instinct MI300A “accelerated processing units” (APU) (CPU–GPU bundle with 128 GB HBM3 each) with 4 APUs per node. El Capitan contains 11424 batch nodes (for a total of 45696 GPUs) [29]. It was deployed in 2024 and delivered 1.742 EFlop/s in its June 2025 Linpack measurement, rising to 1.809 EFlop/s (#1) by November 2025 as commissioning completed, at a power draw of 29.6 MW. It also heads the HPCG November 2025 list.

**Frontier (AMD MI250X) – ORNL, USA** Frontier at Oak Ridge National Laboratory [30] was the first publicly acknowledged exascale system, deployed in 2022. It uses AMD MI250X GPUs (4 per node, 128 GB each), an AMD EPYC “Trento” CPU, and achieves 1.353 EFlop/s (#2) at 24.6 MW, with an HPCG result of 14.05 PFlop/s (#3), on the November 2025 lists. Among its early LQCD scientific results is a study of QCD constraints on isospin-dense matter and the nuclear equation of state [31].

**Aurora (Intel GPU Max) – ANL, USA** Aurora at Argonne National Laboratory [32] uses Intel Xeon Max CPUs (2 per node, 128 GB HBM + 1024 GB DDR5) and Intel Data Center GPU Max (“Ponte Vecchio”) GPUs (6 per node, 128 GB HBM each), achieves 1.012 EFlop/s (#3) at 38.7 MW,

and a 4th place on the HPCG list, both November 2025. It was deployed in 2023. Its unified OneAPI programming model is relevant for LQCD code portability.

### 5.1 JUPITER: the first exascale system in Europe

JUPITER [33] is the first European exascale computer, currently being deployed at the Jülich Supercomputing Centre (JSC) in Germany. It was built by the ParTec/Eviden consortium. The system began phased deployment into production in early 2025, reaching 1 Exaflop in November 2025.

The JUPITER system is designed to implement a Modular Supercomputing Architecture (MSA) [34], with modules based on different complementary architectures to service the diverse computational needs of different workflows, or different steps in a single workflow. The present deployment consists of a storage module and the JUPITER Booster module.

The JUPITER Booster contains approximately 6000 nodes in ~125 racks of BullSequana XH3000, each node carrying 4× NVIDIA Grace-Hopper GH200 superchips. Each GH200 (code-name CG1) contains 72 Arm Neoverse V2 cores (4×128-bit SVE2) with 120 GB LPDDR5 and an H100 GPU (132 SMs, 96 GB HBM3) connected via NVLink-C2C at 900 GB/s. Linpack performance: 1.000 EFlop/s (TOP500 #4, November 2025 list), becoming the first European exascale system and the most energy-efficient of the four exascale systems at 63 GFlops/Watt. FP8 peak: >70 EFlop/s; power: 15.8 MW (Linpack). Note: the June 2025 list measured 0.79 EFlop/s as the system was still being brought fully online; full exascale performance was confirmed at SC25 in November 2025.

The module data center includes the storage subsystem (ExaFLASH: 29 PB NVMe; ExaSTORE: 308 PB HDD; ExaTAPE: 370 PB tape) and the 200/400 Gb/s NVIDIA Mellanox InfiniBand NDR network completes the system. The modular data centre (Eviden) has an area of ~2300 m<sup>2</sup> and includes 7 IT modules (20 racks each), UPS and generator systems, and 15×2.5 MW power stations.

A Universal Cluster module is planned with approximately 1300 nodes in ~14 racks of BullSequana XH3000, each with 2× SiPearl Rhea1 processors. The Rhea1 is a European ARM-based CPU with 80 Neoverse V1 cores and 256-bit SVE, 256 GB DDR5, and 64 GB HBM2e.

Future modules may focus on quantum computing or neuromorphic architectures.

### 5.2 Impact on LQCD

In Figure 7, we see in one plot from [35], the combined impact of algorithm and hardware improvement in recent years on LQCD solver performance. Specifically the authors compare performance on the CPU-only Frontera machine at TACC to some of the machines described in this section and, with the addition of a batching solver algorithm, realize a 240× reduction in required node-hours.

## 6. Hardware trends and the future

### 6.1 AI takes over driving hardware development

The dominant trend shaping current and future HPC procurement is the explosive demand from AI training workloads. As of June 2025, some of the largest GPU clusters in the world

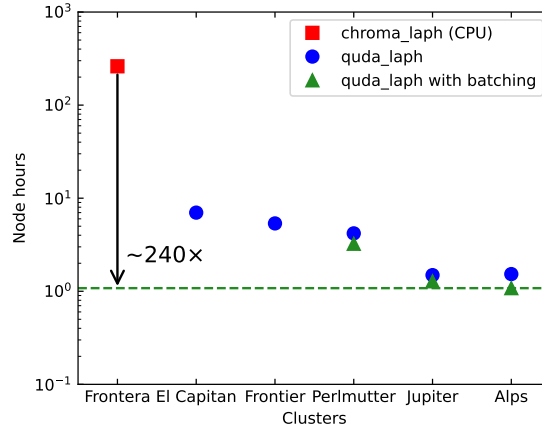


**Figure 7:** Solver speedup resulting from migration from CPUs (Frontera) to modern GPUs, with additional gains from a ‘batching’ solver algorithm (green triangles). [35]

are now operated by technology companies rather than scientific laboratories: the xAI Colossus system in Memphis has 200,000 H100-equivalent GPUs, while Meta, Microsoft/OpenAI, and others operate clusters of 100,000. Project Stargate in the USA represents a \$O(100) billion investment in AI supercomputers, likely dwarfing the aforementioned installations. In comparison, JUPITER’s 23,500 GH200 equivalents represent the largest publicly funded research computer in Europe but are modest on the global AI-infrastructure scale.

On the policy side, the EU has announced €20 billion for large datacentres as part of a broader €200 billion AI fund. The US DOE is also deploying AI-specific systems: Mission and Vision at LANL, Solstice and Equinox at ANL, and Lux at ORNL.

## 6.2 Low-precision hardware

The AI ecosystem’s focus on training-time efficiency and inference has driven GPU vendors to invest massively in low-precision tensor-core hardware, in some devices at the expense of FP64 capability. Table 1 and Table 2 summarise the key specifications of recent and upcoming NVIDIA and AMD devices, respectively. In both cases the ratio of tensor-core FP4 to native FP64 performance grows dramatically.

In other words, vendors are focused on reducing the cost of training models and inference. The energy cost per FLOP drops steeply with lower precision: relative to FP32 FMA (1.0×), FP16 Tensor Core MMA costs 0.12×, FP8 0.06×, and INT8 only 0.04×. At the lowest floating-point precision, FP4-E2M1 allows only the floats:

$$-6, -4, -3, -2, -1.5, -1, -0.5, -0, +0, +0.5, +1, +1.5, +2, +3, +4, +6$$

which is naturally insufficient for native LQCD calculations.

However, there remain opportunities to exploit this low-precision performance and efficiency in GPUs in LQCD calculations. FP64 emulation using INT8 tensor cores has been demonstrated to

<sup>2</sup>Listed here is the GH200 variant with the larger HBM3e memory.

**Table 1:** Key specifications of recent and upcoming NVIDIA GPU devices. Performance numbers in TFlop/s.

| Device             | FP64 TC | TF32 TC | FP4 TC  | Mem (GB)    | BW (TB/s)  | Max TDP (W) |
|--------------------|---------|---------|---------|-------------|------------|-------------|
| H100 PCIe          | 51      | 756     | —       | 80 (HBM2e)  | 2.0        | 350         |
| GH200 <sup>2</sup> | 67      | 989     | —       | 144 (HBM3e) | 4.8        | 1000        |
| B100               | 30      | 900     | 7,000   | 192 (HBM3e) | 8.0        | 700         |
| B200               | 40      | 1100    | 9,000   | 192 (HBM3e) | 7.7        | 1,000       |
| B300               | 1.2     | 1250    | 15,000  | 270 (HBM3e) | 8.0        | 1400        |
| GB200              | 80      | 2200    | 18,000  | 372 (HBM3e) | 16         | 2700        |
| Vera Rubin         | TBD     | TBD     | 100,000 | ≈576 (HBM4) | ≈44 (est.) | yes         |

**Table 2:** Key specifications of recent and upcoming AMD GPU devices. [36] Performance numbers are in TFlop/s

| Device       | FP64 | FP32 TC | FP4 (Matrix) | Mem (GB)    | BW (TB/s) | TDP/TBP (W) |
|--------------|------|---------|--------------|-------------|-----------|-------------|
| MI250X       | 47.9 | 47.9    | —            | 128 (HBM2e) | 3.2       | 560         |
| MI300X       | 81.7 | 163.4   | —            | 192 (HBM3)  | 5.3       | 750         |
| MI300A       | 61.3 | 122.6   | —            | 128 (HBM3)  | 5.3       | 760         |
| MI325X       | 81.7 | 163.4   | —            | 256 (HBM3e) | 6.0       | 1000        |
| MI350X       | 72.1 | 144.2   | 9200         | 288 (HBM3e) | 8.0       | 1000        |
| MI355X       | 78.6 | 157.3   | 10100        | 288 (HBM3e) | 8.0       | 1400        |
| MI450 Series | TBD  | TBD     | 40000        | 432 (HBM4)  | 20        | TBD         |

deliver 2–3× better GFlops/Watt than native FP64 [37]. Mixed precision is not new to LQCD [38] (deflation, inexact solvers), and this trend is likely to intensify.

### 6.3 LQCD performance scaling on NVIDIA hardware

A striking illustration of NVIDIA GPU generational progress for LQCD is provided by the Wilson Dslash benchmark ( Fig. 8) . Originally written in 2008 for the Tesla architecture, the benchmark achieves about 12 TFlop/s on the B200 compared to ~100 GFlop/s on the original Tesla. The improvement of ~100× over 17 years tracks the improvement in hardware closely.

### 6.4 Upcoming traditional HPC systems

The next wave of leadership-class HPC systems foreseen for the scientific community includes:

- **Discovery** [39] (ORNL, USA): AMD EPYC “Venice” + AMD MI430X;
- **Blue Lion** [40] (LRZ, Germany): NVIDIA Vera Rubin;
- **Doudna** [41] (NERSC, USA): NVIDIA Vera Rubin;
- **FugakuNEXT** [42] (RIKEN, Japan): Fujitsu/NVIDIA hybrid.
- **Alice Recoque** [43] (CEA, France): AMD EPYC “Venice” + AMD MI430X;

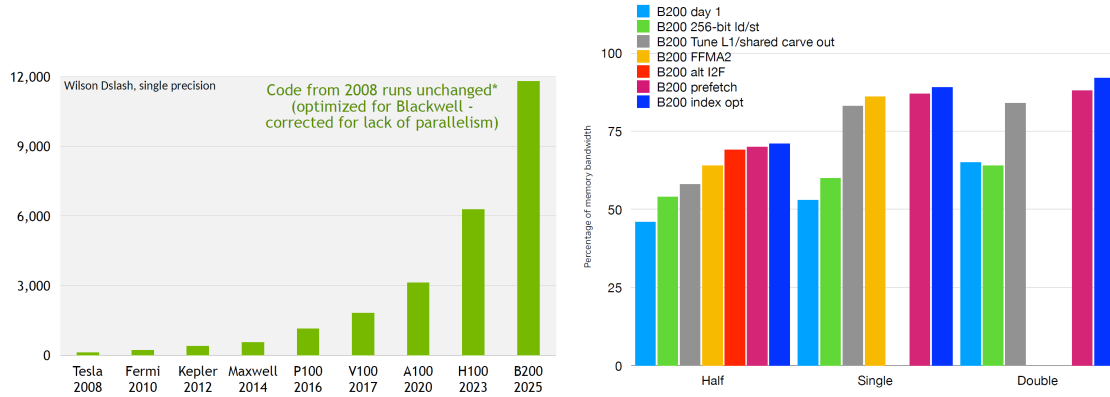


**Figure 8:** Left: Wilson Dslash benchmark (single precision) on NVIDIA GPUs from Tesla (2008) to B200 (2025), using code originally written for the Tesla architecture with only parallelism corrections for Blackwell. Performance grew from ~100 to ~12,000 GFlop/s (~100× over 17 years). Right: percentage of peak memory bandwidth achieved for the Wilson stencil on a single B200 with progressive optimisations. [19].

## 6.5 Usage consequences

Several themes emerge for LQCD software and workflow adaptation. The first is performance portability. Libraries such as Kokkos, that can target multiple architectures without algorithmic changes, become increasingly important. A minapp for the staggered kernel with a Kokkos [44] backend demonstrated good performance across NVIDIA and AMD devices [45]. Second, users should consider task scheduling and job bundling. Exascale nodes are large; schedulers on these machines favour large jobs. Workflow tools (e.g. Flux) and code-level job bundling can improve node utilisation and avoid leaving CPUs idle. Finally, a growing concern is data management. The community needs extensions to the ILDG format to cover new file types and data objects, such as propagators.

## 7. Exotic and emerging hardware

**RISC-V** RISC-V is a free and open instruction-set architecture with growing adoption in both embedded and AI-accelerator markets. Its energy efficiency and open and extensible ISA are promising for HPC use. The European DARE project [46] combines open RISC-V architectures, chiplet technologies, and a co-designed software ecosystem targeting European digital autonomy in HPC and AI.

**Stencil accelerators** The STX stencil accelerator, developed through the EPI, EuPilot, STXDemo, and STXMOD projects [47–49], builds on top of the RISC-V instruction set. The architecture is specifically designed for energy efficiency on stencil-like operations, which are central to LQCD. [50]

**Neuromorphic computing** Neuromorphic computing uses brain-inspired architectures for highly energy-efficient matrix–vector operations. Spiking neural networks (e.g. SpiNNaker [51], Intel Loihi [52]) mimic neuron switching at low precision, while multilevel-cell memristor arrays enable

nanosecond-timescale computation limited only by noise and parasitic currents. Potential applications for stochastic computing and true random number generation are of interest to the LQCD community.

**Analog photonic computing** Photonic processors such as the Lightmatter Passage M1000 [53] and the Q.ANT Native Processing Server (NPS) [54] use photonic circuits for matrix and nonlinear operations in the analog optical domain. Claimed advantages include reduced cooling requirements (increased computational density) and energy efficiency claims of  $\sim 30\times$  relative to digital counterparts. The Q.ANT system is being tested in the HPC context at JSC (Germany). In [55] a photonic multigrid solver with application to LQCD is discussed.

**Field-programmable gate arrays** FPGAs implement custom hardware pipelines with configurable logic gates, enabling high throughput and low-latency data flow for specific algorithms. There have been several applications of FPGAs to LQCD [56] [57]. While FPGA development (HDL, HLS) is more involved than GPU programming, and scaling to large multi-node lattices requires careful network design, FPGAs offer energy efficiency when properly tailored. Memory bandwidth is often the principal bottleneck.

**Quantum computing** Quantum computing is a very active research direction for lattice gauge theories [58, 59]. The hardware landscape is broad and rapidly evolving, and quite diverse: *Quantum annealers*, such as the D-Wave, are optimised for combinatorial optimisation. *Analog quantum simulators* encode quantum systems directly in controllable physical platforms, such as, e.g. neutral-atom arrays (Pasqal, QuEra), though these systems may also support digital gate-model operation. *Universal digital quantum computers* use sequences of discrete quantum gates on qubits realised in a variety of physical substrates, the most common being superconducting circuits (IBM, Google) [60]. Further approaches, such as photonic quantum computing exist as well. Most paradigms and hardware technologies are currently under active exploration for lattice gauge theory applications. For more information, I refer to Zohreh Davoudi's plenary at this conference [61].

## 8. Summary and outlook

We have reviewed the hardware landscape for LQCD at the time of LATTICE 2025. The key messages are:

1. LQCD has been and remains a major consumer of public HPC cycles, and the community runs effectively on all major platforms.
2. Exascale computing has arrived: El Capitan (#1, 1.809 EFlop/s), Frontier (#2, 1.353 EFlop/s), Aurora (#3, 1.012 EFlop/s), and the European JUPITER (#4, 1.000 EFlop/s) are the four exascale systems on the November 2025 TOP500. JUPITER is Europe's first exascale machine and the most energy-efficient of the four (63 GFlops/Watt on HPL). The aggregate performance of all 500 systems now stands at 14.99 EFlop/s.
3. AI applications are reshaping vendor roadmaps. The massive investment in AI training hardware is accelerating low-precision tensor-core development. Mixed precision for LQCD is nothing new, and bootstrapping from low precision to high precision is well established.

4. Energy efficiency will increasingly constrain scientific HPC. With leading AI clusters consuming tens of megawatts, and wafer costs at advanced nodes soaring, energy-efficient hardware may see a renaissance, potentially including specialised architectures. What this means for LQCD software development is not yet clear, but we should monitor these trends closely.
5. Exotic and emerging hardware (RISC-V, stencil accelerators, neuromorphic, photonic, FPGAs, quantum) offers longer-term opportunities; quantum computing is already an active research avenue for lattice gauge theories.

## Acknowledgements

The speaker thanks Christoph Lehner, Peter Boyle, Kate Clark, Mathias Wagner, Steve Gottlieb, Tilo Wettig, Maria Paola Lombardo, Jacob Finkenrath, Bartosz Kostrzewa, Marco Garofalo, Aniket Sen and Leon Hostetler for discussions and for providing material used in this review.

The authors are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through project no. 533766364 (EXC 3701 "Color meets Flavor"), project no. 511713970 (CRC 1639 NuMeriQS), project no. 460248186 (PUNCH4NFDI), by the MKW NRW under funding code NW21-024-A (NRW-FAIR), and by the Helmholtz Association through the AIDAS laboratory and through EXNET-01-07 and EXNET-02-27.

## References

- [1] P.A. Boyle, C. Jung and T. Wettig, *The QCDOC supercomputer: hardware, software, and performance*, 2003.
- [2] M. Kumar, Y. Baransky and M. Denneau, *The gf11 parallel computer*, *Parallel Computing* **19** (1993) 1393.
- [3] F. Butler, H. Chen, J. Sexton, A. Vaccarino and D. Weingarten, *Hadron masses from the valence approximation to lattice QCD*, *Nucl. Phys. B* **430** (1994) 179 [[hep-lat/9405003](#)].
- [4] CP-PACS, JLQCD collaboration, *Computational cost of full QCD simulations experienced by CP-PACS and JLQCD Collaborations*, *Nucl. Phys. B Proc. Suppl.* **106** (2002) 195.
- [5] A.A. Khan, S. Aoki, G. Boyd, R. Burkhalter, S. Ejiri, M. Fukugita et al., *Dynamical quark effects on light quark masses*, *Physical Review Letters* **85** (2000) 4674–4677.
- [6] Y. Aoki, G. Endrődi, Z. Fodor, S.D. Katz and K.K. Szabó, *The order of the quantum chromodynamics transition predicted by the standard model of particle physics*, *Nature* **443** (2006) 675–678.
- [7] E. Follana, Q. Mason, C. Davies, K. Hornbostel, G.P. Lepage, J. Shigemitsu et al., *Highly improved staggered quarks on the lattice with applications to charm physics*, *Phys. Rev. D* **75** (2007) 054502.
- [8] S. Dürr, Z. Fodor, J. Frison, C. Hoelbling, R. Hoffmann, S.D. Katz et al., *Ab initio determination of light hadron masses*, *Science* **322** (2008) 1224–1227.
- [9] S. Borsanyi, S. Durr, Z. Fodor, C. Hoelbling, S.D. Katz, S. Krieg et al., *Ab initio calculation of the neutron-proton mass difference*, *Science* **347** (2015) 1452–1455.
- [10] S. Borsanyi et al., *Leading hadronic contribution to the muon magnetic moment from lattice QCD*, *Nature* **593** (2021) 51 [[2002.12347](#)].

- [11] M.S. Albergo, G. Kanwar and P.E. Shanahan, *Flow-based generative models for Markov-chain Monte Carlo in lattice field theory*, *Physical Review D* **100** (2019) .
- [12] TOP500 Editors, “TOP500 Supercomputer Listings.” <https://top500.org>, 2026.
- [13] Jülich Supercomputing Centre, “Juwels.” <https://www.fz-juelich.de/en/jsc/systems/supercomputers/juwels>, 2022.
- [14] N.E.R.S.C. (NERSC), “Perlmutter system.” <https://www.nersc.gov/what-we-do/computing-for-science/perlmutter>, 2022.
- [15] A. Herten et al., *Application-driven exascale: The JUPITER Benchmark Suite*, in *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*, p. 1–45, IEEE, Nov., 2024, DOI.
- [16] SciDAC, LHPC, UKQCD collaboration, *The Chroma software system for lattice QCD*, *Nucl. Phys. B Proc. Suppl.* **140** (2005) 832 [hep-lat/0409003].
- [17] K. Rupp, “Microprocessor trend data repository.” <https://github.com/karlrupp/microprocessor-trend-data>.
- [18] F. Global, “Supercomputer Fugaku.” <https://global.fujitsu/en-global/technology/research/fugaku>.
- [19] K. Clark, B. Jo, J. Tu, M. Wagner and E. Weinberg, *Lattice QCD on the NVIDIA Grace-Hopper and Blackwell Architectures*, 2025.
- [20] I. Kanamori, T. Aoyama, K. Kanaya, H. Matsufuru and Y. Namekawa, *Bridge++ 2.0: Benchmark results on supercomputer fugaku*, in *PoS(LATTICE2022)*, p. 284, Sissa Medialab, Jan., 2023, DOI.
- [21] “LUMI Documentation.” <https://docs.lumi-supercomputer.eu/hardware/lumig/>.
- [22] C. Alexandrou, S. Bacchio, J. Finkenrath, R. Frezzotti, M. Garofalo, B. Kostrzewa et al., *Status of the ETMC ensemble generation effort*, 2024.
- [23] CINECA, “Leonardo HPC System.” <https://leonardo-supercomputer.cineca.eu/hpc-system/>.
- [24] M. Turisini, M. Cestari and G. Amati, *Leonardo: A pan-european pre-exascale supercomputer for HPC and AI applications*, *Journal of large-scale research facilities JLSRF* **9** (2024) .
- [25] L. Mazur et al., *SIMULATEQCD: A simple multi-GPU lattice code for QCD calculations*, *Computer Physics Communications* **300** (2024) 109164.
- [26] Jülich Supercomputing Centre, “JUWELS Booster overview.” <https://apps.fz-juelich.de/jsc/hps/juwels/booster-overview.html>.
- [27] PUNCH4NFDI project, “Lattice QCD Benchmarks.” <https://results.punch4nfdi.de/?md=/docs/Compute/Benchmarks/lqcd-benchmarks.md>.
- [28] Lawrence Livermore National Laboratory, “El Capitan: NNSA’s first exascale machine.” <https://asc.llnl.gov/exascale/el-capitan>.
- [29] Lawrence Livermore National Laboratory, “El Capitan.” <https://hpc.llnl.gov/hardware/compute-platforms/el-capitan>.
- [30] Oak Ridge National Laboratory, “FRONTIER Direction of Discovery.” <https://www.olcf.ornl.gov/frontier/>.

- [31] R. Abbott, W. Detmold, M. Illa, A. Parreño, R.J. Perry, F. Romero-López et al., *QCD constraints on isospin-dense matter and the nuclear equation of state*, *Physical Review Letters* **134** (2025) 011903.
- [32] Argonne National Laboratory, “Aurora Exascale Supercomputer.” <https://www.anl.gov/aurora>.
- [33] Jülich Supercomputing Centre, “JUPITER the Arrival of Exascale in Europe.” <https://www.fz-juelich.de/en/jsc/jupiter>.
- [34] E. Suarez, N. Eicker and T. Lippert, *Modular supercomputing architecture: From idea to production*, pp. 223–255, CRC Press (2019), DOI.
- [35] M.A. Clark et al., *Using Exascale Computing to Explain the Delicate Balance of Nuclear Forces in the Universe*, **2510.26313**.
- [36] AMD, “AMD Instinct Accelerators.” <https://www.amd.com/en/products/accelerators/instinct.html>.
- [37] H. Ootomo, K. Ozaki and R. Yokota, *DGEMM on Integer Matrix Multiplication Unit*, 2024.
- [38] M. Clark, R. Babich, K. Barros, R. Brower and C. Rebbi, *Solving lattice QCD systems of equations using mixed precision solvers on GPUs*, *Computer Physics Communications* **181** (2010) 1517–1528.
- [39] O.R.N. Laboratory, “Discovery.” <https://www.olcf.ornl.gov/olcf-resources/compute-systems/discovery/>.
- [40] L.S. Centre, “Blue Lion based on Vera Rubin chip architecture.” <https://www.lrz.de/en/news/detail/blue-lion-based-on-vera-rubin-chip-architecture>.
- [41] NERSC, “Doudna System.” <https://www.nersc.gov/what-we-do/computing-for-science/doudna-system>.
- [42] “FugakuNEXT.” <https://www.r-ccs.riken.jp/en/fugaku-next/>.
- [43] CEA, ““EuroHPC JU et le consortium Jules Verne choisissent Eviden pour construire Alice Recoque, le prochain supercalculateur exascale européen”.” <https://www.cea.fr/english/Pages/News/eviden-deliver-alice-recoque-new-european-exascale-supercomputer.aspx>, 2025.
- [44] C. Trott, L. Berger-Vergiat, D. Poliakoff, S. Rajamanickam, D. Lebrun-Grandie, J. Madsen et al., *The kokkos ecosystem: Comprehensive performance portability for high performance computing*, *Computing in Science Engineering* **23** (2021) 10.
- [45] S. Schlepphorst and S. Krieg, *Benchmarking a portable lattice quantum chromodynamics kernel written in kokkos and mpi*, in *Proceedings of the SC '23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis, SC-W '23*, (New York, NY, USA), p. 1027–1037, Association for Computing Machinery, 2023, DOI.
- [46] “DARE project.” <https://dare-riscv.eu/>.
- [47] EPI Consortium, *European Processor Initiative (EPI)*, 2018.
- [48] EuPilot Consortium, *The European PILOT (EuPilot): Pilot using Independent, Local and Open Technologies*, 2021.
- [49] STXDemo, *Proof of Concept for the STX Stencil/Tensor Accelerator*, 2022.

- [50] S. Schlepphorst and S. Krieg, *Porting Lattice QCD benchmark to upcoming STX stencil/tensor accelerator*, in *PoS(LATTICE2024)*, (Trieste, Italy), p. 451 pp., Sissa Medialab Trieste, Italy, 2025, DOI.
- [51] S.B. Furber, F. Galluppi et al., *The SpiNNaker project*, *Proceedings of the IEEE* **102** (2014) 652.
- [52] M. Davies, N. Srinivasa et al., *Loihi: A neuromorphic manycore processor with on-chip learning*, *IEEE Micro* **38** (2018) 82.
- [53] Lightmatter, “Passage™ M1000 3D Photonic Superchip.” <https://lightmatter.co/products/m1000/>, 2025.
- [54] Q.ANT GmbH, “Native Processing Server (NPS): Photonic Computing for Energy-Efficient AI and HPC.” <https://qant.com/photonic-computing/>, 2025.
- [55] T. Lee, F. Brücknerhoff-Plückelmann, J. Dijkstra, J.M. Pawłowski and W. Pernice, *Integrated photonic multigrid solver for partial differential equations*, 2025.
- [56] O. Callanan, A. Nisbet, E. Ozer, J. Sexton and D. Gregg, *Fpga implementation of a lattice quantum chromodynamics algorithm using logarithmic arithmetic*, in *19th IEEE International Parallel and Distributed Processing Symposium*, pp. 8 pp.–, 2005, DOI.
- [57] P. Korcyl and G. Korcyl, *Implementation of the conjugate gradient algorithm in Lattice QCD on FPGA devices*, *PoS LATTICE2018* (2019) 313.
- [58] C.W. Bauer, Z. Davoudi et al., *Quantum simulation for high-energy physics*, *PRX Quantum* **4** (2023) 027001.
- [59] L. Funcke, T. Hartung, K. Jansen and S. Kühn, *Review on Quantum Computing for Lattice Field Theory*, *PoS LATTICE2022* (2023) 228 [2302.00467].
- [60] P. Krantz, M. Kjaergaard et al., *A quantum engineer’s guide to superconducting qubits*, *Applied Physics Reviews* **6** (2019) 021318.
- [61] Z. Davoudi, *Quantum simulation of gauge theories for particle and nuclear physics: Past, today, and future*, in *Lattice 2025*, 2025.