The Role of GPU Computing in the Commercialization and Scale-Up of Fluidized Bed Conversion Processes

Peter Blaser[1], Andrew Larson[1], James Parker[1], Ali Akhavan[1], Niraj Mehta[1]

[1] CPFD Software, 1255 Enclave Parkway, Houston, Texas, 77077, USA

Corresponding Author: Peter Blaser, peter.blaser@cpfd-software.com

Abstract

Simulations of fluidized bed systems are inherently complex, involving multiphase hydrodynamic, thermal, and chemical reaction mechanisms. Additionally, fluidized beds are utilized for a broad range of end-use processes spanning traditional and novel industries and application areas. As such, the role of the particulate phase also varies, with the particles serving diverse purposes including mixing, heat transfer, catalysis, reactant, or product, to name a few. Due to these complexities, simulation has had a smaller impact on the commercialization, scale-up, and troubleshooting of fluidized bed conversion processes than that realized in other chemical engineering applications.

GPU computing, originally used for computer graphics, is at the center of a parallel computing revolution that has occurred over the last decade. Until the 2010s, parallel computing was almost exclusively undertaken using clusters of computers, and later multiple cores on those computers or clusters thereof, collectively called CPU parallelization. Modern GPUs have thousands of compute cores (compared with tens in a CPU), and are rapidly overtaking CPU clusters for HPC applications and artificial intelligence / machine learning (AI/ML) tasks.

This paper explores how advances in GPU and multi-GPU computing have impacted 3D, transient simulations of fluidized bed conversion processes, resulting in simulations running up to 400x faster using GPUs in a single workstation compared with CPU-only performance. While speed enables faster solutions, and marginally faster time-to-market for new technologies, the real benefit is that previously intractable problems are now possible to solve with meaningful resolution (spatial, temporal, physical model complexity, chemical reaction mechanism detail).

Sample case studies using the commercial Barracuda Virtual Reactor® software are shown for applications including the gasification of municipal solid waste streams, chemical looping combustion, and fluidized catalytic cracking. Implications on the breadth of R&D activities, and the acceleration of technology development, commercialization, scale-up, and IP protection are discussed.

Keywords: Fluidized Bed, Computational Fluid Dynamics, GPU Computing, Gasification, Chemical Looping Combustion, Fluidized Catalytic Cracking

1. Background and Motivation

Fluidized beds have been in commercial use for nearly a century since the introduction of the Winkler coal gasification process in the 1920s. Fluidization, a process whereby solid particles are suspended by the drag force exerted from a moving fluid, results in multiple, desirable characteristics for many chemical engineering applications. In particular, fluidized beds of particles typically have excellent mixing and heat transfer characteristics, the particles can be easily moved in a fluid-like manner including motion between multiple reactor vessels, and many processes are sufficiently robust to handle a wide range of particle properties including distributions of particle size, density, or composition.

As a result, fluidization technology is used in many industrial application areas including fluidized catalytic cracking (FCC), cement calcination, polyolefin production, gasification, acrylonitrile manufacture, dehydrogenation processes, and pharmaceutical applications to name a few. Recently, processes involving fluidization are increasingly used for diverse sustainability technologies including advanced recycling of plastics, waste-to-energy/fuels/chemicals, renewable fuels, chemical looping combustion and other decarbonisation applications.

Engineers are increasingly turning to simulation tools to understand, troubleshoot, and optimize fluidized bed conversion processes, but simulations of fluidized bed systems are inherently complex. In addition to the multiphase hydrodynamics that are central to fluidization, most simulations also must capture thermal behavior as well as the effects of chemical reactions. Further, due to the broad range of end-use processes, the role of the particulate phase varies significantly, with the particles serving diverse purposes including mixing, heat transfer, catalysis, reactant, or product, to name a few. Due to these complexities, and the effect of such complexities on computational time, simulation has had a smaller impact on the commercialization, scale-up, and troubleshooting of fluidized bed conversion processes than that realized in other chemical engineering applications.

This paper explores how advances in graphics processing unit (GPU) and multi-GPU computing have impacted 3D, transient simulations of fluidized bed conversion processes, resulting in simulations running up to 400x faster using GPUs in a single workstation compared with CPU-only calculations.

2. Overview of GPU Computing

For decades, computer central processing units (CPUs) became predictably faster due to the consistent doubling of the number of transistors used. By shrinking the transistor size while achieving consistent power density, this meant that CPUs doubled in speed roughly every 1.5 or 2 years. These phenomena are generally referred to as Dennard Scaling and Moore's Law. However, in the 2010s speed-ups declined significantly as thermal limitations became significant. More recently CPUs have increased computational throughput via the use of multiple cores, rather than higher clock speeds and larger transistor counts on a single core.

The traditional approach toward acceleration of complex simulations, such as those based upon computational fluid dynamics (CFD), is to distribute the computational burden across multiple CPUs and/or CPU cores in parallel. Many commercial CFD software products use this CPU-parallel approach, especially those developed prior to the 2010s.

Programmable GPUs from NVIDIA were originally intended for the parallel rendering of computer graphics in the late 1990s. However, due to their very high core count and programmable software stack, NVIDIA GPUs were adapted to scientific computer applications in the 2000s. Today, there is a significant difference between the core counts available in CPUs and GPUs: the most powerful CPUs have anywhere between 24 and 96 cores, whereas GPU core counts now approach 8,000 [1,2,3]. This is why GPUs are rapidly overtaking CPU clusters for HPC applications and AI/ML tasks. The resulting performance differences between CPUs and GPUs over time is shown in Figure 1.

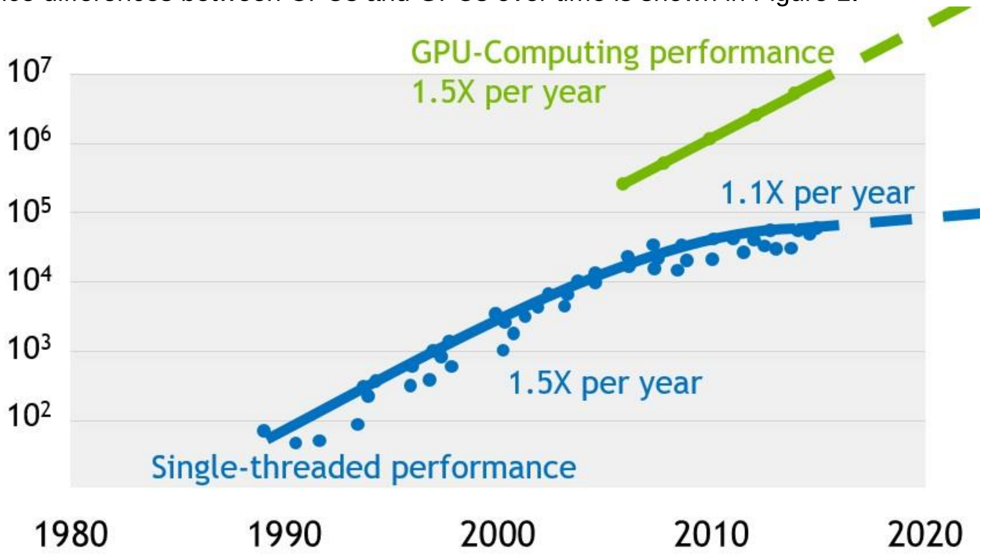

Figure 1. Performance of CPU and GPUs vs time.
Data are shown as circles, lines are curve fits, and dashed lines are projections.

Barracuda Virtual Reactor is a specialized commercial CFD package designed for the simulation of industrial fluid-particle systems, such as fluidized beds, and accelerated via GPU parallelization. The fluid-particle hydrodynamics are solved using an adaptation of the Multiphase Particle-In-Cell (MP-PIC) approach [4-6], an Eulerian-Lagrangian method, to solve the fluid and particle momentum equations in three dimensions with bidirectional coupling between the fluid phase (Eulerian) and the solid particles (Lagrangian). Additional published and proprietary models, including collisional models to capture the effect of nearby particles in more dilute conditions, have been added over the years (e.g. [7]). In the MP-PIC method, a computational particle is defined as a Lagrangian entity in which particles with the same properties such as composition, size, density, and temperature are grouped, allowing industrial-scale systems containing massive numbers of particles to be analyzed using tens of millions of computational particles without losing the advantages of discretizing the solid phase in a Lagrangian frame of reference. The particles thereby retain their discrete nature, with each particle having a unique size, density, composition, temperature, etc. The particle-fluid heat transfer and reaction chemistry are coupled with the particle-fluid hydrodynamics to model the interdependencies between the hydrodynamics, temperature, and composition.

The GPU parallelization of Virtual Reactor™ took place in stages, with an initial release in late 2013. A systemic approach was used whereby the most computationally-intensive functions were ported first, which in 2013 included the major particle functions and linear pressure solver. Continued parallelization, porting of thermal and chemical reaction functions, and related memory management debottlenecking continued through 2020, with the first multi-GPU version released in 2021. Additional details of GPU parallelization can be found elsewhere [8-12].

3. GPU Speed-Up of Fluidized Bed Simulations

Sample GPU and multi-GPU speed-ups are shown in Figure 2 for various bubbling bed and circulating fluidized bed cases studied. Case 1 is a fairly coarse simulation with approximately 250,000 cells and less than 30 million computational particles while Case 4 is a larger model employing just under one million cells and about 55 million particles. Four different applications were tested, and all simulations were run on the same NVIDIA DGX Station using A100 GPUs.

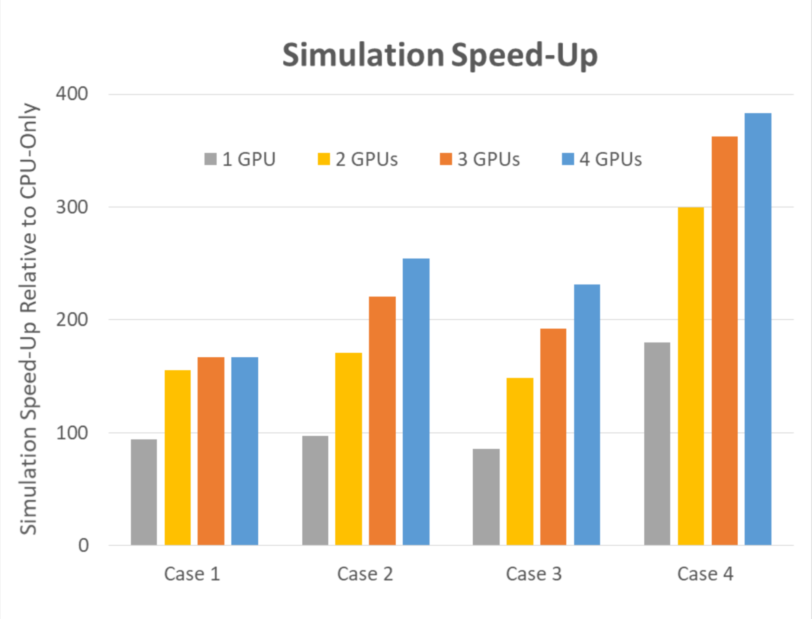

Figure 2. GPU and multi-GPU speed-ups relative to CPU-only for multiple bubbling bed and circulating fluidized bed cases.

In general, speed-ups between 50x and 400x are observed. All cases show additional speed-up when utilizing two GPUs compared with one GPU. Cases 2, 3, and 4 show additional speed-up when using 3 or 4 GPUs. For case 1, little additional speed-up is observed for the 3 and 4 GPU tests due to the smaller model size.

Several factors contribute to simulation acceleration. The aforementioned references [8-12] generally describe enhancements to the algorithms used by the CFD code, which were important contributors to the single GPU performance and enabled the multi-GPU capabilities.

However, another important consideration is the fact that the GPU parallelization is built upon the NVIDIA Compute Unified Device Architecture (CUDA). Thus, as NVIDIA continues to innovate and develop GPU hardware and provide enhancements to the CUDA framework and software stack, the resulting benefits immediately impact new versions of Virtual Reactor as well as software users running on the latest hardware. This effect can be seen in Figure 3 which shows simulation speed for single and dual GPU simulations of Case 2 from Figure 2. Figure 3 shows the A100 GPU performs over 6x faster than the older cards in single-GPU mode, and nearly 5x faster in dual-GPU mode. These results were obtained using the same software version with the only difference being the various GPU cards used, which were released over roughly a four year span.

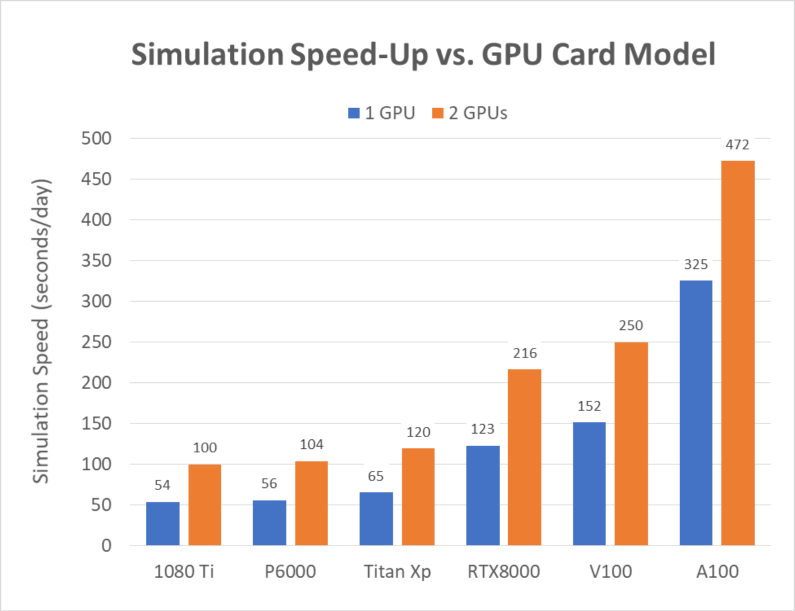

Figure 3. GPU and multi-GPU speed-ups vs. GPU card model.

4. Impact of GPU Computing on Fluidized Bed Simulations

4.1 Faster Results and Increased Number of Cases Simulated

On the surface, the obvious impact of GPU computing on fluidized bed simulations is that results are obtained much faster than was possible previously. For example, a 150-200x speed-up means that a week-long simulation is completed in an hour. Or, put another way, 150-200 permutations of a model could be run sequentially in the time that previously was required to run a single case.

The ability to simulate two orders of magnitude more cases in the same time period has a significant impact on the commercialization and scale-up of fluidized bed conversion processes. Typically simulation is used to study a broad range of possibilities during the research and development phase, prior to cold-flow or pilot testing. Additionally, with the down selection afforded by extensive CFD testing, only the top candidate designs are further evaluated with physical testing. As a result, new technologies are brought to market faster and at a significantly lower total cost [13], and simulation results are used to communicate benefits of the technology with customers, partners, investors, and in support of patent applications.

To quantify sample speed-up, consider the simulation of a chemical looping combustion (CLC) reactor located at the National Energy Technology Laboratory (NETL) in Morgantown, West Virginia. This model was originally developed by CPFD Software and NETL in 2012 [14] and has previously been used to demonstrate later CFD improvements [15,11]. A schematic of the CLC model is shown in Figure 4.

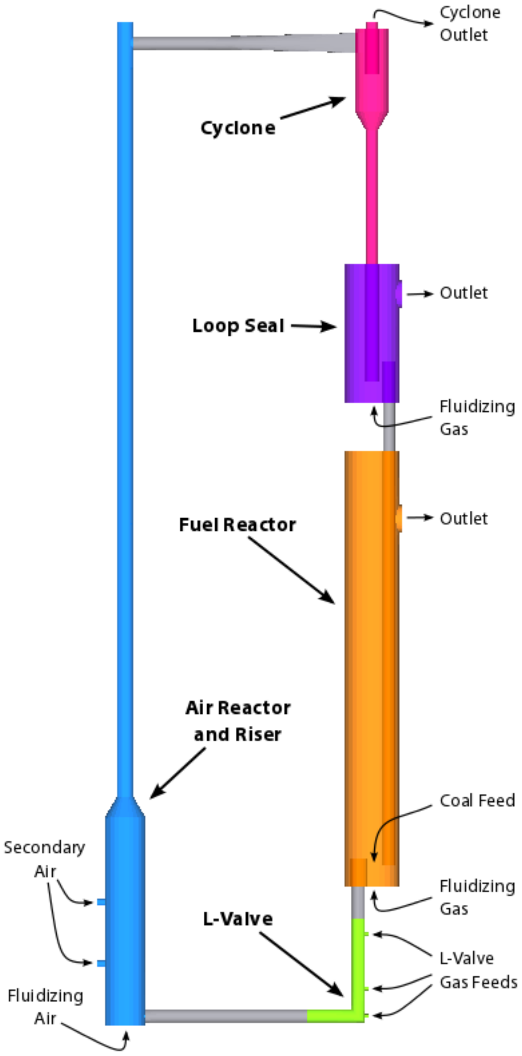

Figure 4. Diagram of Chemical Looping Combustion Reactor Model

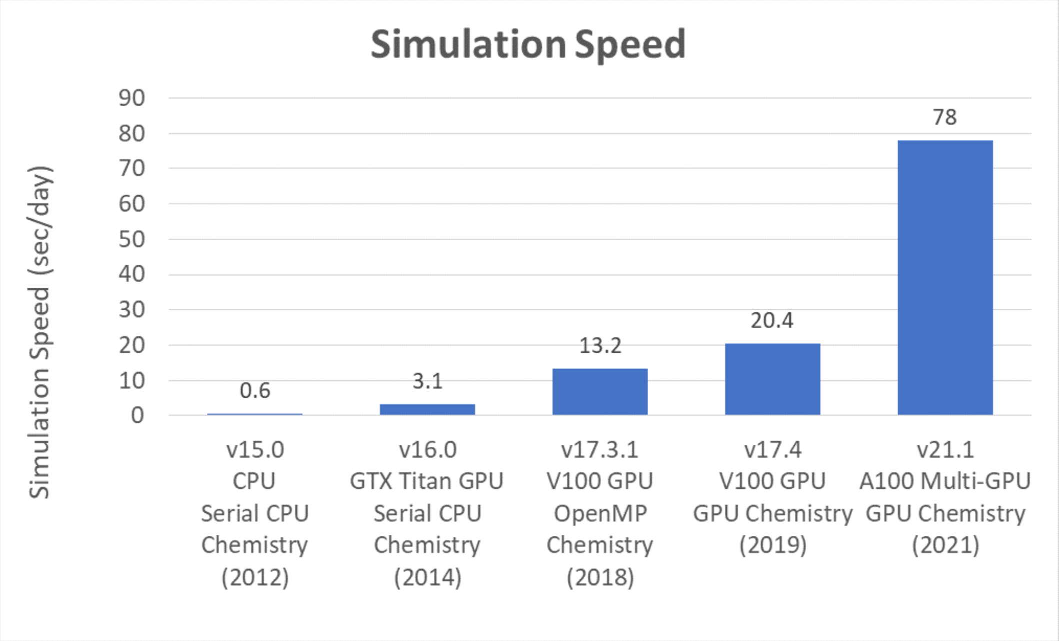

Figure 5. Maximum calculation rates of CLC model by release and then-current hardware capabilities.

Because of the complexity of the CLC model – it is a three-dimensional model of a full circulation loop where gas-particle hydrodynamics, heat transfer, and reaction chemistry are simulated – and the previous history of contemporaneous reporting of model setup and calculation rates, the CLC model is a good benchmark model for demonstrating current CFD capabilities and studying the recent history of model calculation rates. Figure 5 shows the maximum calculation rates of the CLC model by release using the hardware available at that time. The simulation speed increased from 0.6 seconds per day on a single CPU in 2012 to 78 seconds per day in 2022 – a speedup of 130x. This is on the lower-end of the speed-ups reported herein due to the relatively small model size; a model created a decade ago is too small to utilize all the GPU cores available today. Larger speed-ups are observed for larger models.

4.2 Higher Fidelity Models

In fact, few engineers would create the same models today as would have been created a decade ago. In practice, the increased computational bandwidth affords modelers the opportunity to create higher fidelity models. This is typically manifested in larger computational domains, finer spatial resolution (finer grids), including more physical models (radiation, evaporation, liquid film transfer, collisional details), or utilizing more complex chemical reaction mechanisms (species, reactions, rates).

To illustrate this, consider the model of an FCC regenerator originally created in 2012, as shown on the left of Figure 6. This example represents the first 3D MP-PIC regenerator simulation which included gas-catalyst hydrodynamics, heat transfer and reaction kinetics associated with coke combustion and emissions [16]. The model was created to assist in identification of the root cause of afterburn observed at the refinery, whereby the dilute phase gas temperature at the top of the regenerator was nearly 100°F (56°C) higher than that measured in the dense phase of the fluidized bed.

Prior to model formulation, several questions were raised regarding factors which could have contributed toward the afterburn problem. Perhaps the spent catalyst was skewed to one side at the top of the spent catalyst riser? And, regardless of whether the riser outlet flow was uniform or not, what impact did the 17 arms of the spent catalyst distributor have on uniformity of catalyst injection into the dense bed? Or, was the afterburn a function of the regenerator design and layout of internals?

Since it was not feasible then to run the lift line, spent catalyst distributor, and regenerator with internals in a single model, the domain was subdivided into three component models as shown in Figure 6. Model 1 was used to determine the fluxes of gas and particles at the top of the lift line. A non-uniformity was observed, and the gas and solid flux profiles were used as an input to Model 2, which found that the spent catalyst distributor performed fairly well at overcoming the non-uniformity. The subsequent flow distribution from Model 2 was then used as boundary conditions into Model 3, which successfully predicted the magnitude and radial nonuniformity of the afterburn observed at the refinery.

Model 1 was only used to determine gas-catalyst hydrodynamics and thus was simplified with isothermal and non-reacting assumptions. This small model ran reasonably fast, with 150 seconds of realtime simulated in approximately two days on a single core Intel i7 CPU processor. Model 1 contained about 30,000 cells and 600,000 computational particles.

It was then hoped that Model 2 could be used to complete the analysis. Model 2 was a thermal, reacting model of the spent catalyst distributor and regenerator, containing about 400,000 cells and 4.6 million computational particles. However, after six weeks little progress was made and the simulation speed was slowing. It was later determined that 250-300 seconds were required to reach a quasi-steady state for that particular model, which would have taken approximately three years to complete in 2012!

Thus, Model 3 was created, using the information obtained from Model 2 as boundary conditions. Model 3 used 240,000 cells and 2.6 million computational particles, reaching 300 seconds in just over two months.

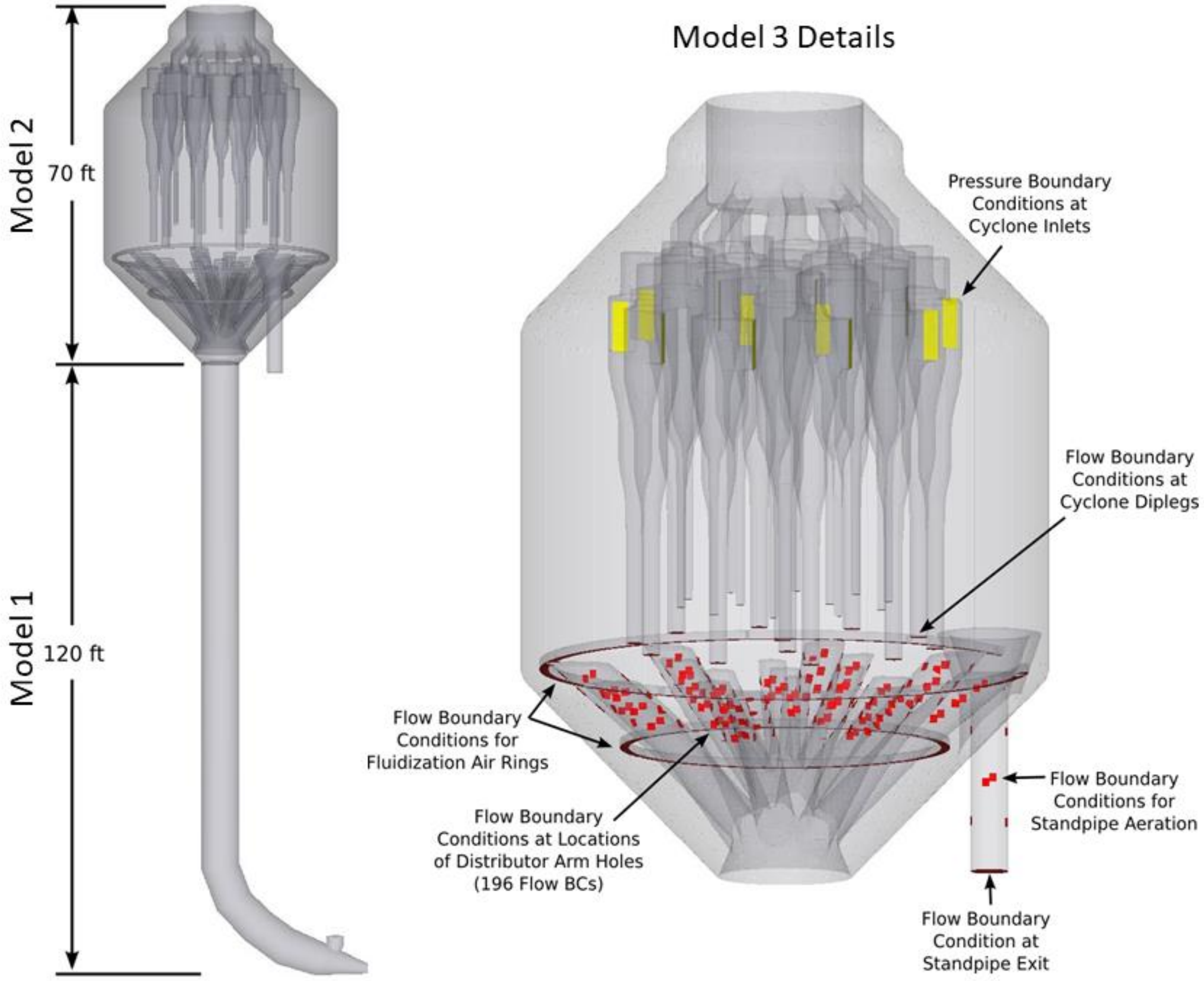
Model 2
70 ft
Model 1
120 ft
Model 3 Details
Pressure Boundary Conditions at Cyclone Inlets
Flow Boundary Conditions at Cyclone Diplegs
Flow Boundary Conditions for Fluidization Air Rings
Flow Boundary Conditions at Locations of Distributor Arm Holes (196 Flow BCs)
Flow Boundary Conditions for Standpipe Aeration
Flow Boundary Condition at Standpipe Exit

Figure 6. FCC Regenerator Simulation with Multi-Model Formulation

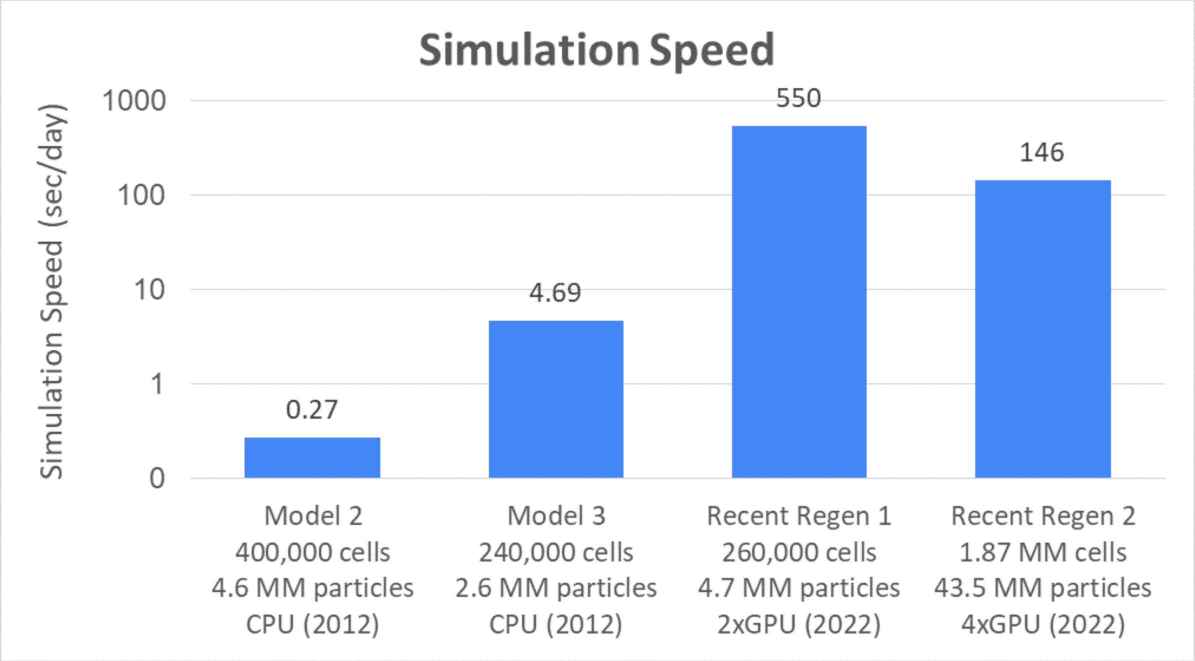

Figure 7. FCC Regenerator Simulation Speed with Multi-GPU Parallelization (log scale)

While the original 2012 model is no longer available for comparison, a timing test was performed on a recent FCC regenerator project. The simulation had 260,000 cells and 4.7 million computational particles and thus was roughly comparable to the prior models. Figure 7 shows a log plot which compares the simulation speed for the current regenerator of 550 seconds per day using two A100 GPUs against the 2012 cases. 550 seconds per day means that the 300 second end time could now be achieved in approximately 13 hours. This is about two thousand times faster than Model 2, and over one hundred times faster than Model 3 from 2012.

A higher fidelity model was also benchmarked. The right most data point in Figure 7 is from a model with 1.87 million cells and 43.5 million computational particles. With a multi-GPU simulation speed of 146 seconds/day, a higher resolution 300 second simulation is now possible in about two days.

4.3 Previously intractable problems are now possible

The cumulative outcome of faster simulation times, increased numbers of parametric simulations, and the newfound capability to run much higher fidelity models, is that previously intractable problems are now made possible, and can be solved in industrially-relevant time scales. As an example, consider how Thermochem Recovery International, USA (TRI) leveraged a digital syngas model to develop and commercialize a gasification process for the conversion of municipal solid wastes (MSW) to jet fuel. CFD simulation was used for all fluidized beds used in process development including cold flow testing and reacting cases at test, process demonstration, and commercial scale. The project received a Highly Commended designation at the IChemE Global Awards 2020, with additional project details reported elsewhere [17].

Early work on the project was first reported in 2009, and a 70x speed-up was observed over the following decade, primarily due to single GPU acceleration [18]. The subsequent multi-GPU development, together with other hardware and software enhancements, was shown to increase the speed-up from 70x to 1,500x through 2022 [12], which is consistent with other results reported herein. However, it is important to note that the types of simulations performed in 2009 and the early 2010s did not include all the complexity of the most recent models. In fact, many of today's models were not possible to run a decade ago, whether due to physical memory limitations or the timescales involved.

And this is the point. Waste-to-fuels and other sustainability and decarbonisation applications such as the chemical recycling of plastics, are changing our planet for the better. But the complexity involved in these types of processes – multiple particle species, widely ranging sizes and compositions, liquid film and droplet collisional transfer, vaporization, solid particle chemistry, devolatilization, gas phase chemistry, and so forth – is something that is only now possible to compute in a meaningful, engineering time scale. Fortunately the advent of GPU and multi-GPU computing has happened concurrently with the timely need to solve some of our planet's toughest challenges.

5. Accessibility of GPU Computing Resources

GPU computing resources are widely available today and eliminate the need for a costly CPU cluster. A sample case study showed that a single computer built on NVIDIA GPU technology has the same computational capacity as a cluster of 300 CPU servers, but requires only 1/18th of the power consumption at 1/8th of the cost, all within a workstation computer [3].

On-premise GPU hardware typically falls in one of two categories. First, GPUs are already present in most workstation computers, and usually more can be added. Many GPU-accelerated fluidized bed models are run on workstation computers. However, for higher-end computing needs, specialized multi-GPU systems are often used. For example, NVIDIA DGX systems contain 4-8 GPUs in either a workstation or rack-mounted footprint and deliver exceptional performance [19]. As of the time of writing, most GPU and multi-GPU systems could be acquired in the $10,000 - $150,000 USD price range.

In recent years GPU computing resources have also become accessible on major cloud service providers [20-22]. The cloud provides CFD users instant access to virtually-unlimited high performance computing resources without the overhead of purchasing and maintaining on-premise hardware.

6. Conclusions

GPU computing has dramatically expanded the possibilities for 3D, transient simulations of fluidized bed conversion processes. Simulations were observed to run between 50 and 400 times faster using GPU parallelization compared with CPU performance on the same system.

Several impacts of speed-up have been explored. Practically-speaking, the reduction in simulation time means that engineers can now simulate two orders of magnitude more cases in the same time as was previously required for a single model. This enables a broader virtual exploration of possibilities during research & development and more targeted physical testing. As a result, new technologies are brought to market faster and at a significantly lower total cost

With faster simulations, modelers can now formulate higher fidelity models. This is typically manifested in larger computational domains, finer spatial resolution, including more physical models, or utilizing more complex chemical reaction mechanisms. Such higher fidelity models are needed to address some of our planet's toughest sustainability challenges, and many of these models were simply not possible to run a decade ago. Fortunately the advent of GPU and multi-GPU computing makes the previously-impossible practical, and does so using hardware which is readily available today.

References

[1] For CPU core count examples see https://www.techradar.com/news/intel-and-amd-face-an-armed-onslaught-from-a-96-core-cpu-monster and https://en.wikipedia.org/wiki/Fugaku_(supercomputer). Fugaku was listed at #1 in the Top 50 supercomputers and uses 48 core CPUs.
[2] For GPU core counts see NVIDIA A100 Tensor Core GPU Architecture whitepaper available from https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf.
[3] For an overview of Moore's Law, Dennard Scaling, and the increases in CPU and GPU speed over time see "NVIDIA GPUs: What and Why" as presented by Reynaldo Gomez at the Barracuda Virtual Reactor Users' Conference 2019: https://cpfd-software.com/users-conference-2019-nvidia-gpus-what-and-why/ Figure 1 taken from slide 2. Computational comparison with CPU clusters based upon the comparison between the NVIDIA DGX-2 and a cluster of 300 Broadwell CPU servers.
[4] Andrews M, O'Rourke P, The multiphase particle-in-cell (MP-PIC) method for dense phase particulate flows, International Journal of Multiphase Flow. 22, (1996) 379-402.
[5] Snider D, O'Rourke P, Andrews J, Sediment flow in inclined vessels calculated using multiphase particle-in-cell model for dense particle flows, International Journal of Multiphase Flow 24, (1998), 1359-1382.
[6] Snider D, An incompressible three-dimensional multiphase particle-in-cell model for dense particle flows, Journal of Computational Physics, 170, (2001), 523-549.
[7] O'Rourke PJ, Zhao P, Snider D, A model for collisional exchange in gas/liquid/solids fluidized beds, Chemical Engineering Science, 64, (2009), 1784-1799.
[8] Larson A, Quickly Applying GPU Acceleration to Barracuda VR: a MP-PIC CAE Software, NVIDIA NVIDIA GTC 14, 2014. PDF is accessible at https://on-demand.gputechconf.com/gtc/2014/presentations/S4417-gpu-acceleration-barracuda-mp-pic-cae-software.pdf
[9] Larson A, Carver T, Zhao P, CUDA Accelerated Lagrangian Interpolation to Cartesian Grid, Eleventh International Conference on CFD in the Minerals and process Industries, 7-9 December 2015, CSIRO, Melbourne, Australia, PDF is accessible at https://www.cfd.com.au/cfd_conf15/PDFs/089LAR.pdf.
[10] Larson A, Parker J, User-Defined Drag Models on the GPU, NVIDIA GTC Conference Poster P6107, 2016. PDF is accessible at https://on-demand.gputechconf.com/gtc/2016/posters/GTC_2016_Computational_Fluid_Dynamics_CFD_08_P6107_WEB.pdf
[11] Parker J, Larson A, Application of Recent CFD Advancements to the Modeling of Chemical Looping Systems, 5th International Conference on Chemical Looping, 24-27 September 2018, Park City, Utah, USA.
[12] Larson A, Blaser P, Waste-to-Fuels Technology Enables by GPU and Multi-GPU Simulation, NVIDIA GTC 21, April 12-16, 2021. Video playback of the presentation at this virtual conference is available here: https://cpfd-software.com/gtc_21_presentation_playback/

[13] Singh R, Computational modelling in FCC design and troubleshooting, Barracuda Virtual Reactor Users' Conference 2019.
[14] Parker J, Simulation of Coal Particles in a Full Chemical Looping Combustion System. Final Report submitted to US Department of Energy, National Energy Technology Laboratory, March 2012.
[15] Parker J, Thibault S, Williams K, Advancements in the CFD Modelling of Clean Coal Technologies, Clearwater Clean Coal Conference 2014, Clearwater, FL.
[16] Clark S, Snider D, Fletcher R, Multiphase Simulation of a Commercial Fluidized Catalytic Cracking Regenerator, AIChE Annual Meeting, Pittsburgh, 2012.
[17] Blaser P, Chandran R, Digital Technology Enables Novel Waste-to-Fuels Process, IChemE Global Awards 2020 Presentation. Playback available at https://cpfd-software.com/cpfd-and-tri-highly-commended-at-icheme-global-awards-2020/.
[18] Chandran R, Advances in Barracuda Simulations of TRI Steam Reformer, Barracuda virtual Reactor Users' Conference 2019. Playback available at https://cpfd-software.com/users-conference-2019-advances-in-barracuda-simulations-of-tri-steam-reformer/.
[19] Gomez R, Lin A, Powering Multi-GPU Simulations on NVIDA DGX™ Station A100, 2021. Webinar playback available at https://cpfd-software.com/powering-multi-gpu-simulations-on-nvidia-dgx-station-a100/.
[20] See FLSmidth Drives Sustainable Productivity while Reducing Time and Costs Using HPC on AWS, available at https://www.hpcwire.com/solution_content/aws/flsmidth-drives-sustainable-productivity-while-reducing-time-and-costs-using-hpc-on-aws/
[21] See FLSmidth transcends on-premises limitations with Azure high-performance computing, available at https://customers.microsoft.com/en-us/story/1417831248134599964-flsmidth-manufacturing-azure
[22] See Barracuda Virtual Reactor on Azure, available at https://techcommunity.microsoft.com/t5/azure-compute-blog/barracuda-virtual-reactor-on-azure/ba-p/3054715

Note, all website references to websites refer to them as accessed on February 17, 2022.