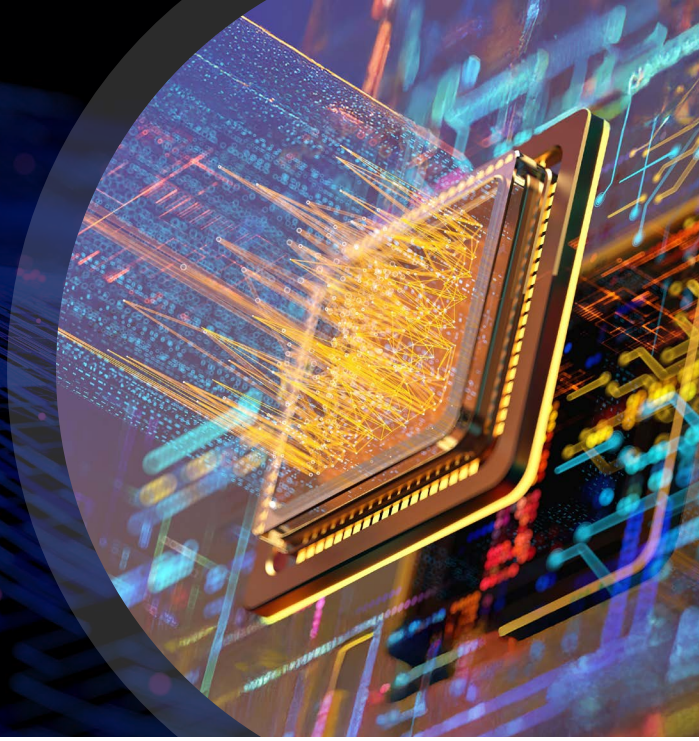


High-performance actuarial modeling

Unlocking the optimal choice with CPUs, GPUs and looped modules

The insurance industry faces growing demands for faster, more cost-effective actuarial modeling. With a new choice of central processing unit (CPU) and graphic processing unit (GPU) technology, supported by looped modules functionality, FIS® Insurance Risk Suite – Prophet is set to give firms the powerful tools they need to optimize runtime and cost.

What are the advantages of adding GPUs to your mix of computational power options?

Our recent work with three clients shows that:

- CPU optimization alone can deliver dramatic improvements.
- GPU performance and cost benefits can be dramatic, but vary by model type and configuration.
- To get the best from GPUs, models may need to be rebuilt with GPUs in mind and not just converted from CPU technology.

GPUs excel at executing large volumes of parallel, independent tasks, such as policy level projections or stochastic simulations, which are common in actuarial modeling. However, actuarial models can often extend beyond linear projections and introduce memory-intensive workloads for the model infrastructure. So, as current GPUs operate with less memory capacity than CPUs, they must be managed carefully to unlock their full potential.

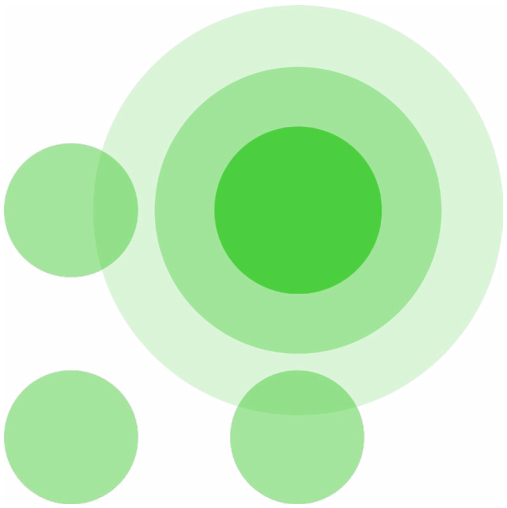
In practice, this means models need to balance memory workloads available for parallelization, often by experimenting with batch sizes and allocation of computing resources. And that's where looped modules come in.

The benefits of the looped modules method

Looped module implementations restructure data into vectors, enabling highly parallelized computing instructions and often delivering performance gains on their own. For Insurance Risk Suite – Prophet, they also provide a scalable framework to leverage vectorized CPU – with optional advanced vector extensions (AVX) – or GPU hardware.

The results of three use cases

- **Case 1:** Critical illness portfolio – CPU optimization cut runtime by 85% and compute cost by 95%; GPUs offered limited benefit due to focus on minimizing rework costs and batch size constraints.
- **Case 2:** Large annuity book – GPUs achieved 60-75% improvement over CPUs for scenario-based loops.
- **Case 3:** Unit-linked product with guarantees – GPUs reduced runtime and compute cost by almost 97%, due to high simulation counts.



The details

Case 1: Critical illness portfolio

A large global insurer faced challenges with a large, complex critical illness portfolio involving 60 products and 900,000 model points, running deterministic liability and stochastic simulations under IFRS 17. Initial runs were slow, batch sizes were small and memory was intensive and costly to run.

What we did:

- Implemented looped modules for both liability and cohort-level models.
- Optimized loops: liability model grouped policies and cohort model looped over simulations.
- Tested across CPUs, AVX and GPUs.

Results:

- CPU optimization with AVX: Reduced runtime by 85% and compute cost by 95%.
- GPU performance: A hybrid CPU/GPU approach improved runtime, but overall, it was still slower due to added preparation time.
- Recommendation: Roll out looped modules broadly on CPUs and explore options to run larger batches and better leverage the power of GPUs.

Case 2: Annuity portfolio for Solvency II

Another large global insurer wanted to assess GPU viability for running accurate annuity models at scale, to replace proxy models for Solvency II.

What we did:

- Tested looped modules with scenario-based loops on GPUs.
- Worked with existing looped modules setups and selected the variation with the largest module to maximize GPU utilization.
- Configured loops over scenarios and tested on five products containing 232,000 model points for 2,500 scenarios.
- Benchmarked across different batch sizes and two GPU machine types (T4 and A100).

Results:

- On the same machine, GPU runs delivered approximately 60-75% runtime and compute cost improvement over CPUs, depending on the chip type.
- We identified future enhancements: faster loop-level result reading, improved parallelism and performance improvements for high-spec GPU cards.

Case 3: Unit-linked product with guarantees

An insurer was exploring GPU efficiency for a stochastic unit-linked product requiring 140,000 model points and 32,000 simulations.

What we did:

- Implemented looped modules using simulation as the loop dimension to leverage GPU efficiency.
- Separated initial model components into data transfers and calculations, with calculations forming the module.
- Restructured stochastic table reads for efficiency.
- Benchmarked on A100 GPU (cut-down run: 10,000 points, 16,000 simulations).

Results:

- Both runtime and compute cost improved by approximately 97%.
- GPU benefit was high, given high simulation count and policy-level stochastic calculations, although memory constraints required careful batch sizing.
- T4 GPUs were cost-effective for smaller runs but hit memory limits on full dataset, whereas A100 handled the full-scale runs without any memory issues.

High-performance computing is transforming actuarial modeling, and both CPU and GPU technologies offer powerful ways to accelerate results. While GPUs can deliver exceptional speed for certain workloads, modern CPUs remain highly efficient and often require less upfront effort to optimize.

For many insurers, CPU-based solutions can achieve strong performance without major architectural changes or extensive rebuilds. The key is to start with what delivers the fastest, most cost-effective gains for your business. Once CPU optimization is complete, you can assess whether GPU acceleration provides enough additional benefit to justify the investment.

As FIS advances GPU capabilities within Insurance Risk Suite – Prophet, we have a unique advantage: our benchmark is our own highly optimized CPU architecture. Years of refinement have given us deep insight into how CPUs (with or without AVX) deliver exceptional performance, and where GPUs can unlock further gains. This hybrid expertise enables us to evaluate your models holistically and recommend the most effective hardware for your models' specific needs. FIS actuaries can help you evaluate both paths and design a strategy that maximizes performance while minimizing complexity. Speak to an expert today.

Unlock more

What are the key takeaways?

1. CPU optimization matters:

Looped modules alone can slash runtime and compute cost dramatically.

2. GPU benefits depend on model design:

High simulation counts and large loops favor GPU.

3. Plan for GPU from the start:

Straight conversions may not always unlock the full potential; some models may need to be designed with GPU architecture in mind.