GPU-Accelerating Process Simulation Performance using NVIDIA's Direct Sparse Linear Systems Solver

Ian Washington
Jeff Renfro
Honeywell

NVIDIA GTC
San Jose, CA
20-Mar-2024

Anton Anders
Kirill Voronin
NVIDIA

AGENDA

- Introduction
- Honeywell Overview
- How is Modeling used in the Process Industries?
- Process Simulation Computational Challenges
- cuDSS Overview
- Performance Study
- Conclusions
- Questions

Copyright © 2024 by Honeywell International Inc. All rights reserved.

INTRODUCTION

Process models are widely used for the simulation of industrial processes, but they can be large and computationally expensive to solve

Honeywell has an ongoing objective to seek performance improvements for our process simulation software

NVIDIA has a new direct sparse linear system solver cuDSS (CUDA Direct Sparse Solver) that uses GPUs for computation acceleration

Honeywell has a collaboration with NVIDIA to provide input on cuDSS features and has done performance testing using cuDSS on problems in the process simulation domain

Copyright © 2024 by Honeywell International Inc. All rights reserved.

HONEYWELL OVERVIEW

NASDAQ: **HON** | ~**715** sites | ~**97,000** employees | **Charlotte, NC** headquarters | **Fortune 500** | 2022 Revenue: **~$35 B**

AEROSPACE TECHNOLOGIES

BUILDING AUTOMATION

ENERGY & SUSTAINABILITY SOLUTIONS

INDUSTRIAL AUTOMATION

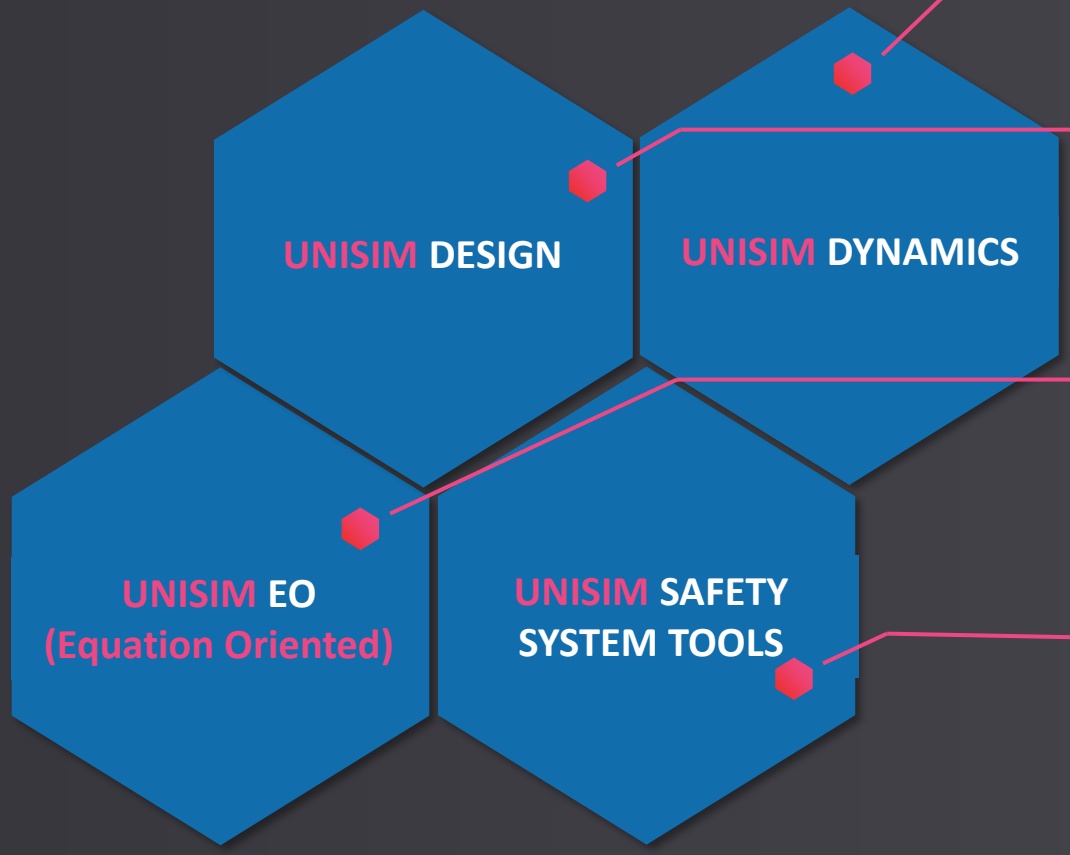
HONEYWELL CONNECTED ENTERPRISE

Shaping the Future Across Industries

Copyright © 2024 by Honeywell International Inc. All rights reserved.

UNISIM DESIGN

Honeywell's Process Modeling and Simulation Product



PROVIDES DYNAMIC SIMULATION CAPABILITY FULLY INTEGRATED WITH UNISIM DESIGN

THE CORE STEADY STATE FLOWSHEETING ENVIRONMENT

HIGH-PERFORMANCE SIMULATION PLATFORM, ENABLING SIMULATION AND OPTIMIZATION IN THE SAME ENVIRONMENT

DESIGN, RATE & TEST SCENARIOS FOR PRESSURE RELIEF, BLOWDOWN & FLARE

https://www.honeywellforge.ai/us/en/products/industrial-operations/honeywell-unisim-design-suite

Copyright © 2024 by Honeywell International Inc. All rights reserved.

HOW ARE MODELS USED IN THE PROCESS INDUSTRIES?

Copyright © 2024 by Honeywell International Inc. All rights reserved.

PROCESS INDUSTRY CHALLENGES

UNPLANNED DOWNTIME

LOST PRODUCTION

NET ZERO GOALS

ENERGY & EMISSIONS

SUPPLY/DEMAND IMBALANCE

MARGIN IMPACT

SKILL LOSS

HUMAN CAPITAL CHALLENGES

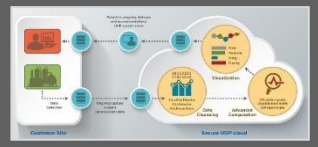
Model Based Applications Are a Part of the Solution To These Challenges

Copyright © 2024 by Honeywell International Inc. All rights reserved.

UNISIM DESIGN: LINKS TO PERFORMANCE SOLUTIONS

Honeywell Process Digital Twin

Monitor, Predict, and Improve Plant Performance

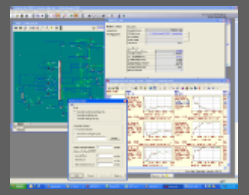


Honeywell Operator Competency Management

Process Operator Training Simulator

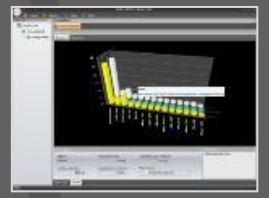
Honeywell Advanced Process Control

Direct link to Process Controller



Corrosion Predict® Suite

Corrosion Management

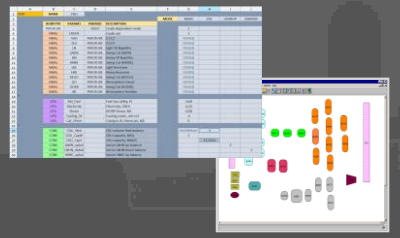


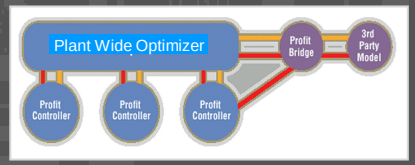
Honeywell Asset Performance Management

Asset Performance Calculation

PrincepsLP®

Supply Chain Optimization





Plant Wide Optimizer

Real Time Closed Loop Optimization

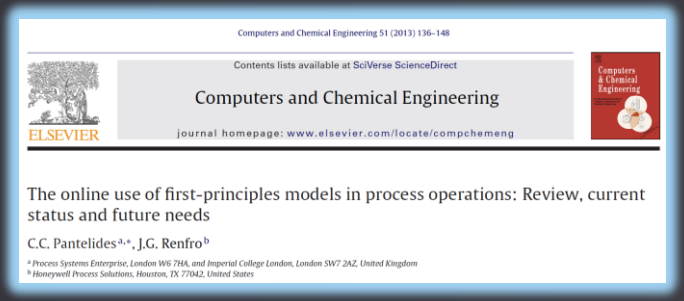
PROCESS DIGITAL TWINS

WHAT IS IT?

Process Digital Twin (PDT) is a live, virtual process representation (process model) of a physical plant

WHAT IS THE VALUE?

PDT enables optimization, emissions and performance monitoring, early issue detection and rapid analysis, including what-if queries


Computers and Chemical Engineering 51 (2013) 136–148

Contents lists available at SciVerse ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

The online use of first-principles models in process operations: Review, current status and future needs

C.C. Pantelides[a,*], J.G. Renfro[b]

[a] Process Systems Enterprise, London W6 7HA, and Imperial College London, London SW7 2AZ, United Kingdom
[b] Honeywell Process Solutions, Houston, TX 77042, United States

HOW DOES IT WORK?

Streams live data into the model for calibration, predicts future operations. Machine learning models can be combined with the first principles model to form a hybrid model with superior prediction quality

Copyright © 2024 by Honeywell International Inc. All rights reserved.

UNISIM DESIGN SUSTAINABILITY MODELING

Carbon Capture and Storage (CCS)

- CO2 Capture - solvent based and membrane separation
- Sustainability Thermodynamics
- CO2 Freeze Out utility

CO2e Emissions Accounting

- Tool auto calculates scope 1 & scope 2 emissions
- Complete component mass balances

Green H2 Production

- Proton Exchange Membrane electrolyzer (PEM)
- Alkaline electrolyzer (AEL)

Copyright © 2024 by Honeywell International Inc. All rights reserved.

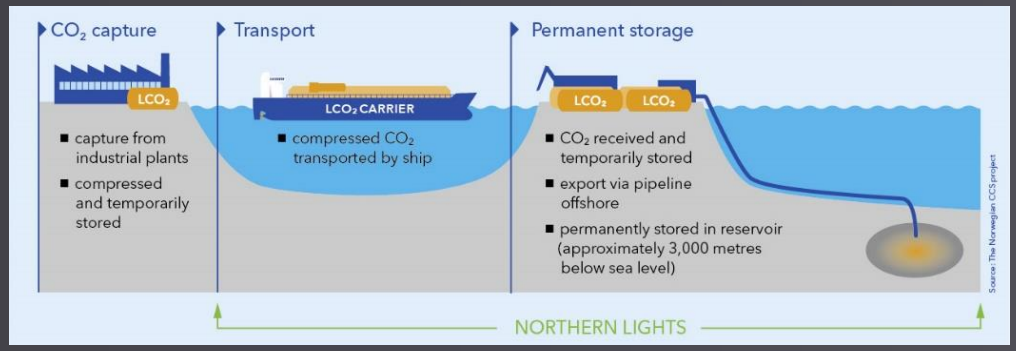
UNISIM DESIGN SUSTAINABILITY MODELING

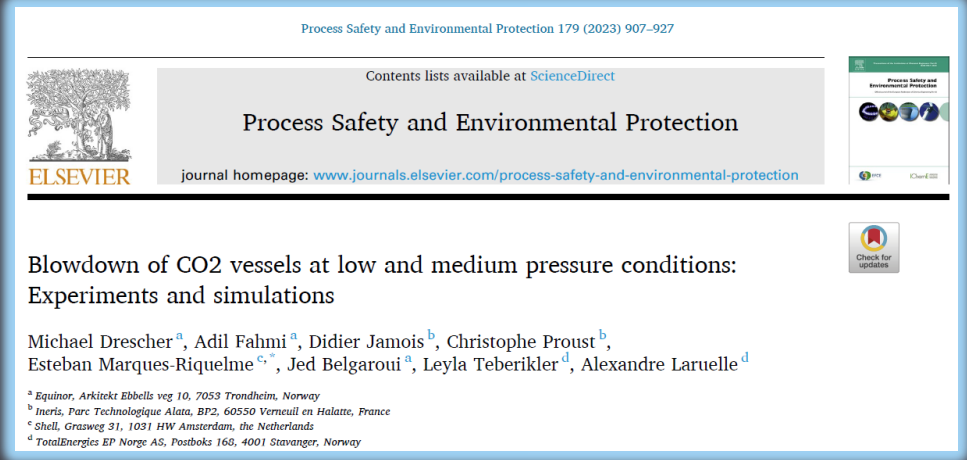
Model CO2 Storage and Removal from Vessels

The UniSim Design EO Blowdown Utility can reliably simulate CO2 depressurization

Detailed nonequilibrium dynamic vessel model predicts key process variables

Recent and past publications show the model has a good match to measured experimental process data





Process Safety and Environmental Protection 179 (2023) 907–927

Contents lists available at ScienceDirect

Process Safety and Environmental Protection

journal homepage: www.journals.elsevier.com/process-safety-and-environmental-protection

ELSEVIER

Blowdown of CO2 vessels at low and medium pressure conditions: Experiments and simulations

Michael Drescher [a], Adil Fahmi [a], Didier Jamois [b], Christophe Proust [b], Esteban Marques-Riquelme [c,*], Jed Belgaroui [a], Leyla Teberikler [d], Alexandre Laruelle [d]

[a] Equinor, Arkitekt Ebbells veg 10, 7053 Trondheim, Norway
[b] Ineris, Parc Technologique Alata, BP2, 60550 Verneuil en Halatte, France
[c] Shell, Grasweg 31, 1031 HW Amsterdam, the Netherlands
[d] TotalEnergies EP Norge AS, Postboks 168, 4001 Stavanger, Norway

Copyright © 2024 by Honeywell International Inc. All rights reserved.

PROCESS SIMULATION COMPUTATIONAL CHALLENGES

Copyright © 2024 by Honeywell International Inc. All rights reserved.

SOLVING LARGE SCALE PROCESS MODELS

Process models contain many equipment models that have sets of variables $\mathbf{x}_i$ and equations $\mathbf{g}_i(\mathbf{x}_i)$ to model the i^{th} unit operation

- Mass, Energy and Momentum Balance Equations
- Phase Equilibrium Equations
- Chemical Reaction Equations

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \mathbf{g}_1(\mathbf{x}_1) \\ \mathbf{g}_2(\mathbf{x}_2) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{g}_{n_u}(\mathbf{x}_{n_u}) \\ \mathbf{g}_c(\mathbf{x}) \end{bmatrix} = \mathbf{0} \qquad x = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_{n_u} \end{bmatrix}$$

The combined process model equation set $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ can become large when it includes a significant amount of process equipment within its scope and/or when the number of chemical components is large and/or when PDE discretization is involved

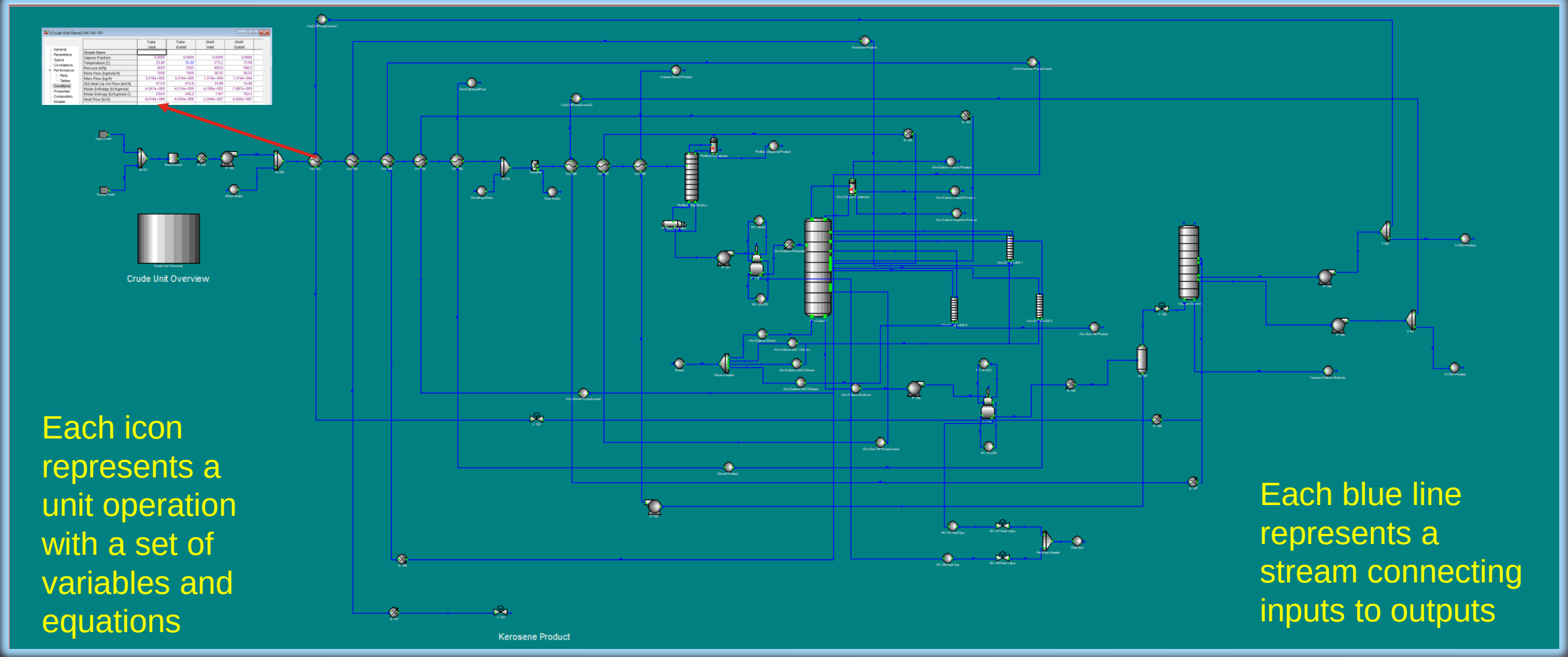
Traditional solution methods rely on sequential approaches that try to decompose the process model equations into smaller solvable subsets

The sequential methods don't perform well for complex models with a lot of interacting variables that can result from heat integration, control system specifications and recycles in process models

Copyright © 2024 by Honeywell International Inc. All rights reserved.

PROCESS FLOWSHEET MODEL EXAMPLE

Atmospheric and Vacuum Crude Unit Model with Preheat Train



Each icon represents a unit operation with a set of variables and equations

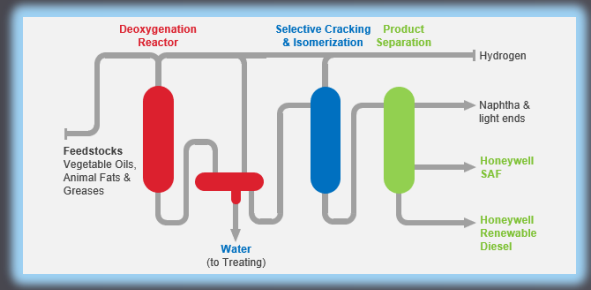
Each blue line represents a stream connecting inputs to outputs

Copyright © 2024 by Honeywell International Inc. All rights reserved.

SOLVING LARGE SCALE PROCESS MODELS

Solving all equations simultaneously overcomes issues of highly interacting variables and allows many other benefits

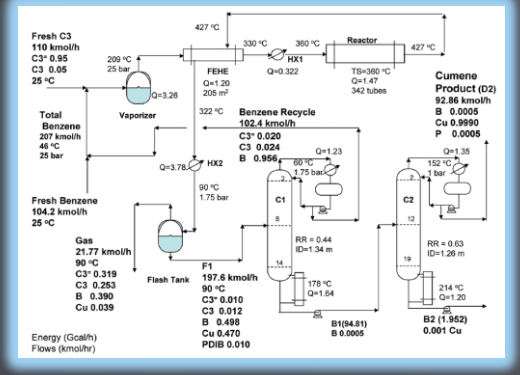
- Specification flexibility
- Allows efficient solution by optimization solvers



Honeywell UOP Two Stage Ecofining Process

Solving all process model equations simultaneously can create a large problem and requires a robust and efficient large-scale nonlinear equation solver

The nonlinear equation solver must solve a large, sparse linear system of equations at each iteration to generate variable changes, which can take more than 90% of the solution time



Design and Control of the Cumene Process, Luyben (2010)

Copyright © 2024 by Honeywell International Inc. All rights reserved.

WHY IS A DIRECT SPARSE SOLVER REQUIRED?

Jacobian matrix is very large and sparse

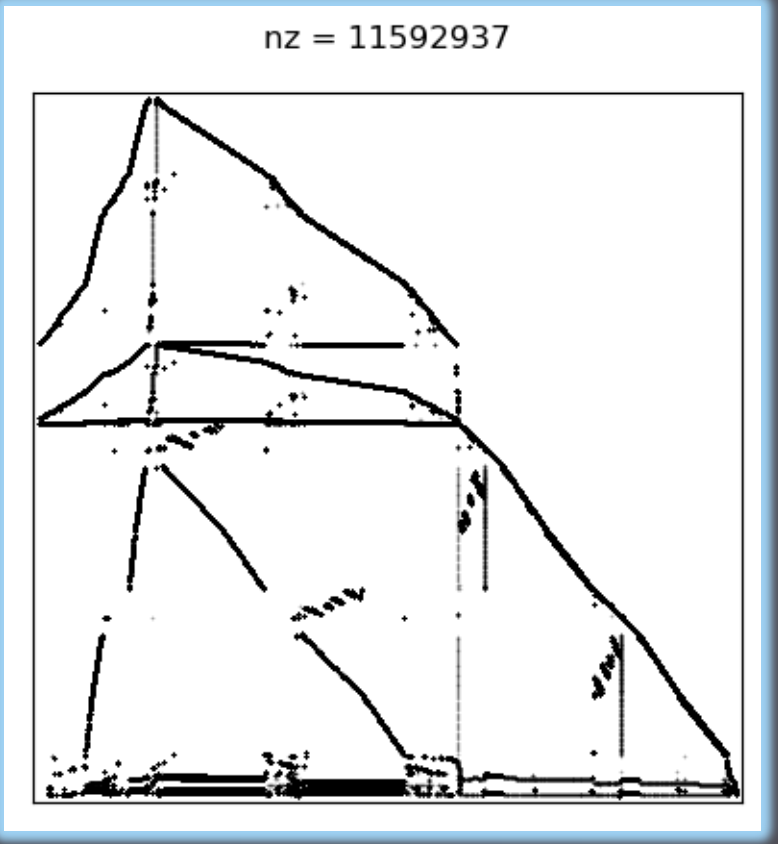
- A dense matrix with $n = 10^6$ requires 8 terabytes of memory in double precision (impractical)

Jacobian matrix is unsymmetric

Indirect methods can be slow and may not be sufficiently stable

Reliability is needed due to significant changes in both variable and Jacobian values throughout the nonlinear solver iterations

Jacobian Matrix Structure



Need a Fast Direct Solver to Meet these Requirements

Copyright © 2024 by Honeywell International Inc. All rights reserved.

UNISIM EO NONLINEAR EQUATION SOLVER

UniSim EO Simulation Executive

NOVA Nonlinear Equation Solver

Process Model $\mathbf{f(x) = 0}$

AXB Sparse Linear Algebra Interface

3rd Party Solver

3rd Party Solver

3rd Party Solver

DOTAXB Solver

cuDSS Solver

Copyright © 2024 by Honeywell International Inc. All rights reserved.

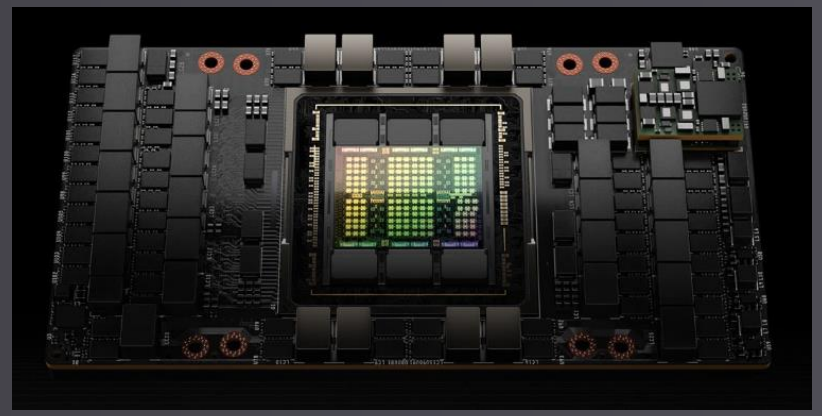
BENEFITS OF SOLVING MODELS FASTER

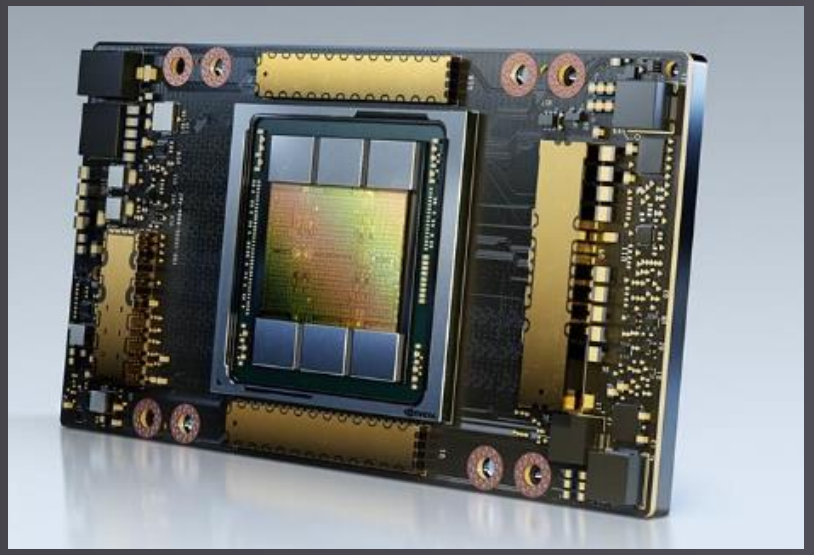
Enables larger scope first principles models to be solved in a reasonable amount of time for reliable process digital twin execution

Eliminates the need for including surrogate model development in an application workflow to achieve faster solutions which reduces maintenance

Improves engineering productivity by being able to complete complex simulations faster

Improves engineering designs by enabling more scenarios to be considered in the same time frame





Copyright © 2024 by Honeywell International Inc. All rights reserved.

NVIDIA / HONEYWELL COLLABORATION

Honeywell established a collaboration with NVIDIA in a meeting in 2020

NVIDIA had begun some initial R&D work on a direct sparse solver that was compatible with Honeywell solver requirements

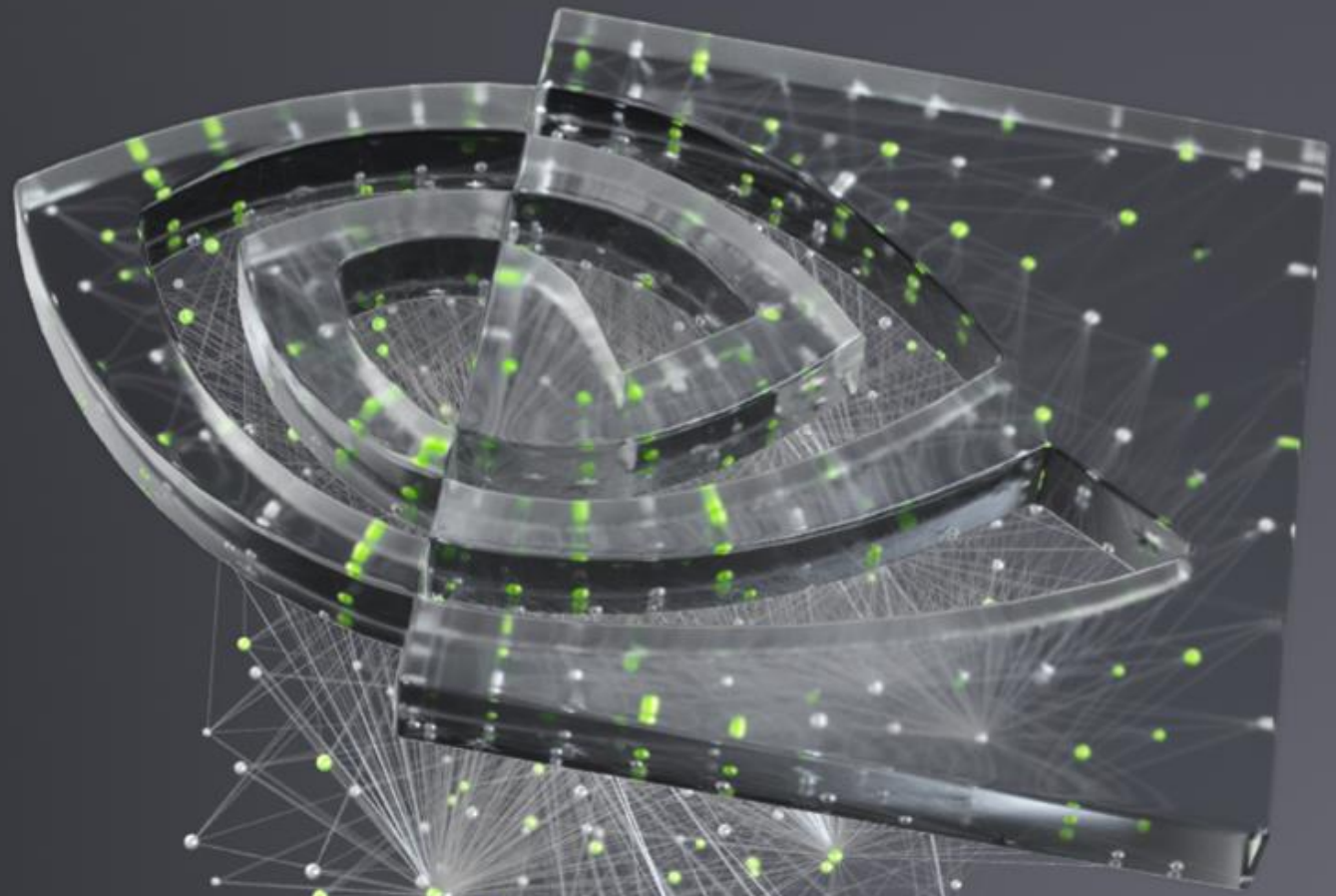
Honeywell

Honeywell submitted test matrix cases to NVIDIA and created an interface to the NVIDIA solver for internal testing

NVIDIA

Honeywell resumed performance testing in 2023 after NVIDIA was ready with a version with plans for productization

Copyright © 2024 by Honeywell International Inc. All rights reserved.



CUDA DIRECT SPARSE SOLVER GTC MARCH 2024

Anton Anders, 03/20/24

CUDA DIRECT SPARSE SOLVER

High-level Overview

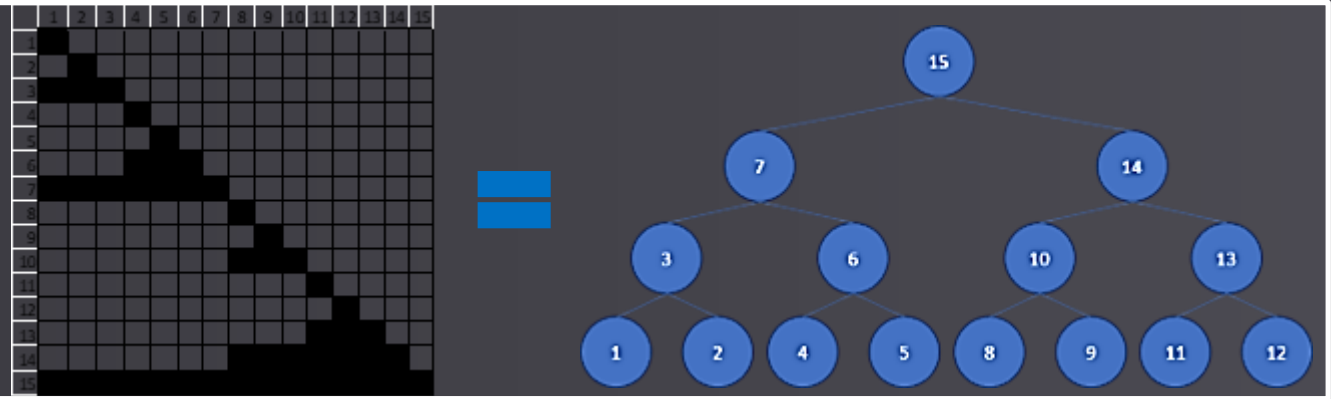
cuDSS	• Solves sparse linear systems on NVIDIA GPU
Used in many scientific domains	• SLAM • Robotics and self driving • circuit simulation • CAE
Robust alternative to iterative solvers	• Can solve ill-conditioned matrices • Does not require any additional information
Ideal fit for GPU architecture	• Utilizes both high memory bandwidth and compute power of GPU
Aims to enhance applications alongside existing solutions:	• CPU: MKL PARDISO, MUMPS, USI PARDISO, SuperLU, KLU • GPU: CHOLMOD, WSMP, GLU

CUDA DIRECT SPARSE SOLVER

Algorithm overview

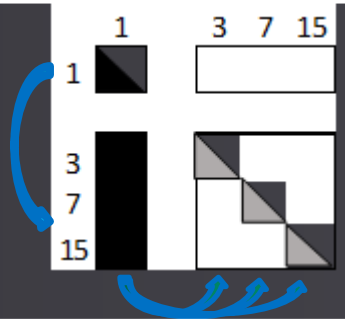
Phase 1: Reordering

Representing initial matrix as a binary graph to extract more parallelism and reduce memory requirements.



Phase 2: Factorization

Factorizing reordered matrix as A = LU (lower triangular x upper triangular matrices)



Simple factorization scheme that consists of 3 main workloads:

1) factorization of the diagonal block
2) TRSM for sub-diagonal blocks
3) GEMM to update next blocks

Phase 3: Solving

Solving the equivalent system with lower and upper triangular factors

Perform forward and backward substitutions with triangular matrices L and U. Optionally it can include iterative refinement process using factorized matrix as a preconditioner.

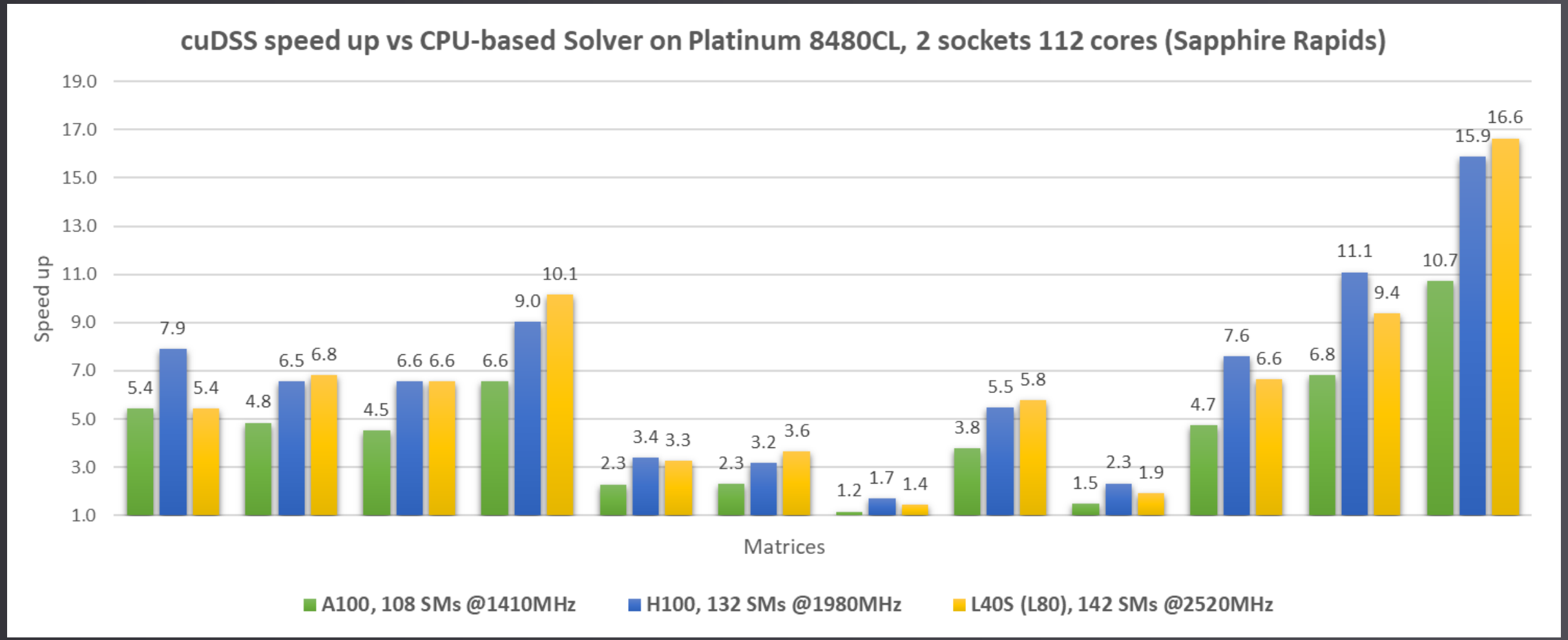
CUDA DIRECT SPARSE SOLVER

Features list

- All matrix types: symmetric (Lower or Upper), Hermitian, positive/non-positive definite, general
- All factorization types: LU, LDL^t, LL^t
- All native data types: double, float, complex, double complex
- Sorted and unsorted CSR (compress sparse row) format for input matrix
- Different reordering schemes: Metis ND, AMD, COLAMD, BTF
- Pivoting strategies: no pivoting, local/global , row/column, pivot threshold, pivot epsilon
- Optional returning of inertia, number of pivots, factors diagonal, perm
- Refactorization
- Optional iterative refinement after solving step
- Multiple right-hand sides on solving step
- Fully asynchronous factorization and solving steps
- Linux and Windows, X86 and ARM

CUDA DIRECT SPARSE SOLVER

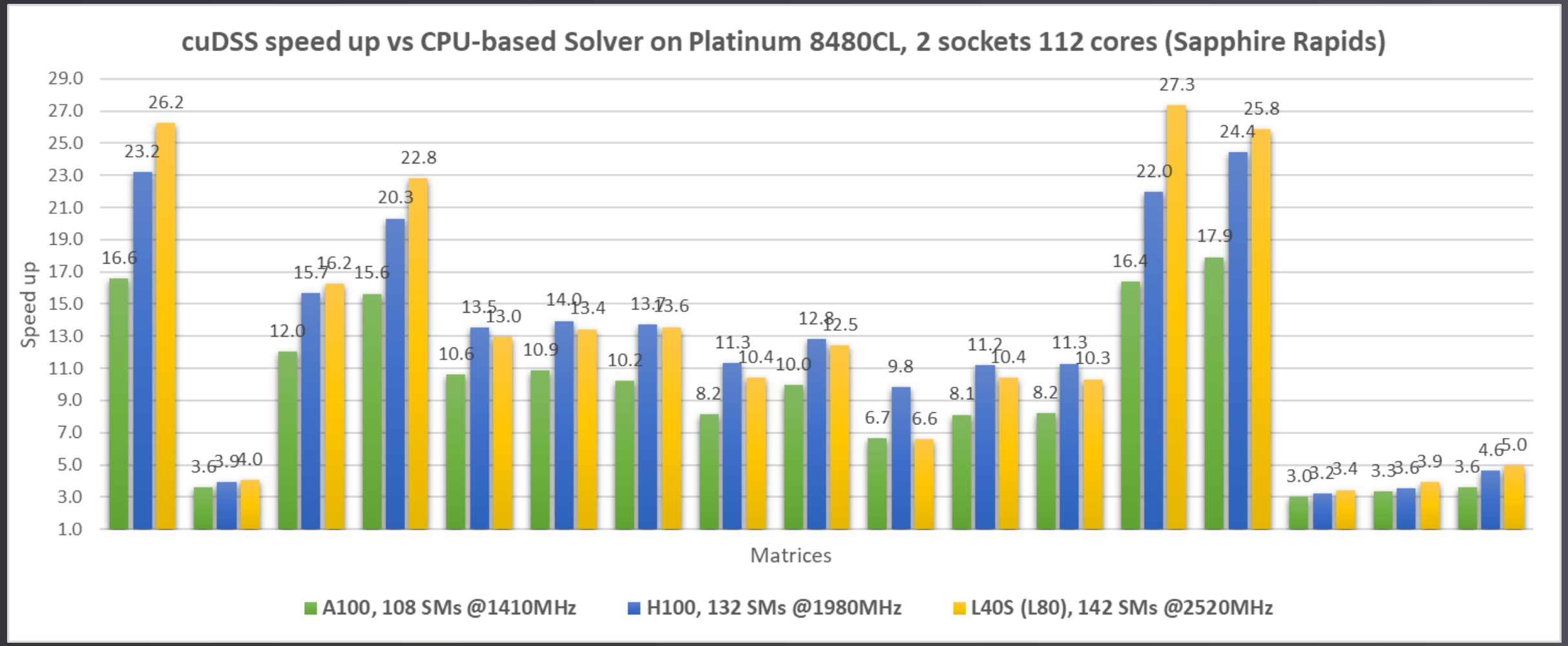
Performance overview



- Non-symmetric matrices from circuit simulation (double precision)
- N from 8K to 4.5M, NNZ from 400K to 20M

CUDA DIRECT SPARSE SOLVER

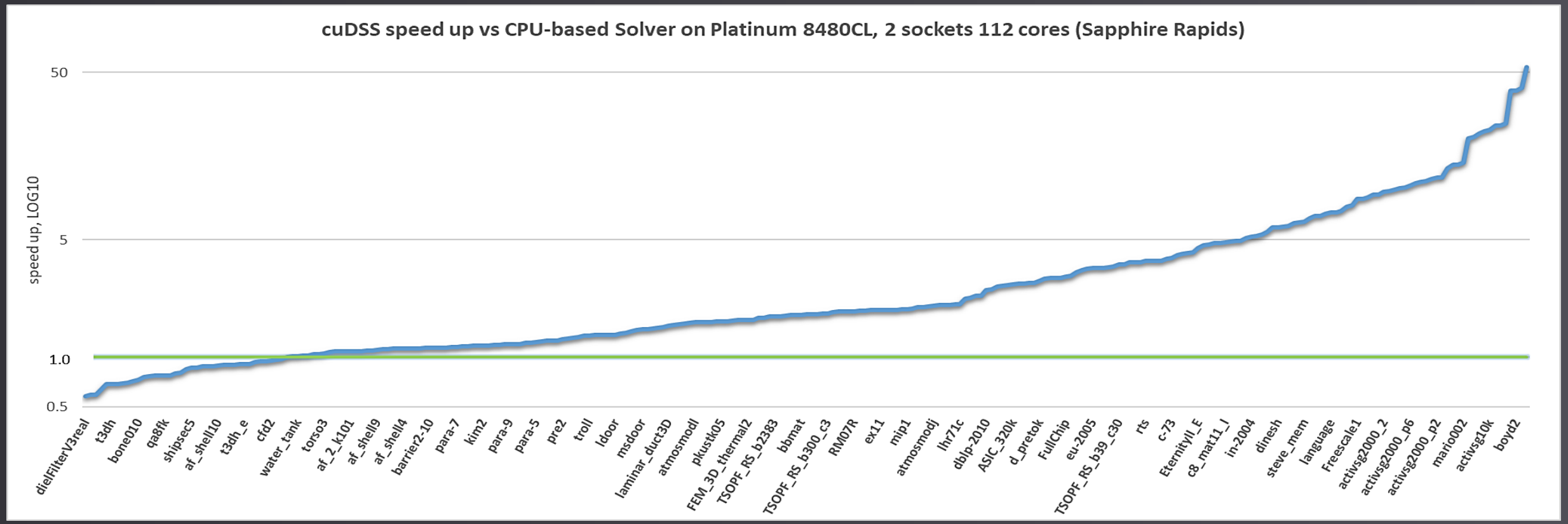
Performance overview



- Symmetric matrices from SLAM (double precision)
- N from 2K to 1.6M, NNZ from 6K to 5M

CUDA DIRECT SPARSE SOLVER

Performance overview



- cuDSS perf on H100, 132SMs @ 1980MHz
- CPU-based Solver on Intel Xeon Platinum 8480CL, 2 sockets 112 cores (Sapphire Rapids)
- 273 symmetric and non-symmetric matrices from Florida Collection (double precision)
- N from 5K to 4.6M, NNZ from 500K to 45M
- Performance summary
  - Reordering: Geomean = 1.2; Speed up: MAX = 5.5, MIN = 0.36; cuDSS faster in 53% of cases
  - Factorization: Geomean = 1.9; Speed up: MAX = 70, MIN = 0.52; cuDSS faster in 76% of cases
  - Solve: Geomean = 2.7; Speed up: MAX = 45, MIN = 0.41; cuDSS faster in 92% of cases

PERFORMANCE STUDY

Copyright © 2024 by Honeywell International Inc. All rights reserved.

LINEAR SYSTEM SOLUTION PERFORMANCE

Reducing linear system solution times yields performance improvements for the overall process model solution

Honeywell has a suite of test matrices in the industry standard *Matrix Market Format*, which are the Jacobian matrices $\mathbf{J}$ from a linearization of the model equations $\mathbf{f}(\mathbf{x}) = \mathbf{0}$

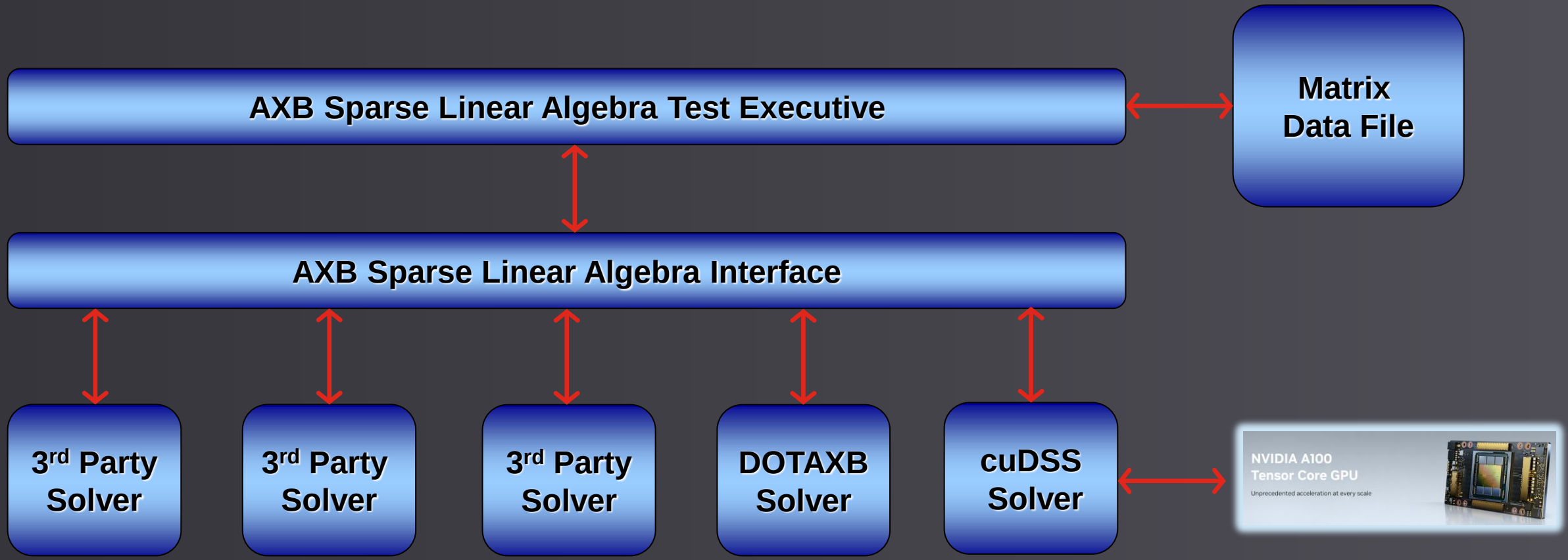
$$\mathbf{J}(\mathbf{x}_k)\Delta\mathbf{x}_k = -\mathbf{f}(\mathbf{x}_k)$$

Basic Newton direction equations

The test matrices are generated from linear systems in large scale upstream/midstream oil & gas, refining, petrochemical and chemical process flowsheet models

Copyright © 2024 by Honeywell International Inc. All rights reserved.

AXB SPARSE LINEAR ALGEBRA TEST EXECUTIVE



Copyright © 2024 by Honeywell International Inc. All rights reserved.

HARDWARE AND SOFTWARE TESTING BASIS

- Microsoft Azure Server
- Windows 10 OS (64 bit)
- CUDA Toolkit 12.3
- NVIDIA DLL: cudss.dll v0.1.0
- UniSim DLL: dotaxb.dll R500
- GPU: NVIDIA **A100** 80 Gb PCIe
- CPU: AMD EPYC 7V13 64-Core Processor (2.44 GHz)
- All test matrices were pre-processed with scaling and numerical zeros below a drop tolerance removed

Size	vCPU	Memory (GiB)	Temp Disk (GiB)	NVMe Disks	GPU	GPU Memory (GiB)	Max data disks	Max uncached disk throughput (IOPS / MBps)	Max NICs/network bandwidth (MBps)
Standard_NC24ads_A100_v4	24	220	64	960 GB	1	80	12	30000/1000	2/20,000
Standard_NC48ads_A100_v4	48	440	128	2x960 GB	2	160	24	60000/2000	4/40,000
Standard_NC96ads_A100_v4	96	880	256	4x960 GB	4	320	32	120000/4000	8/80,000

1 GPU = one A100 card

Copyright © 2024 by Honeywell International Inc. All rights reserved.

cuDSS PERFORMANCE SPEEDUP: COLD START*

Matrix	n	nnz(A)	CUDSS (secs)	DOTAXB (secs)	SpeedUp (DOTAXB/CUDSS)
lgcmpdis	1,136,993	76,789,656	18.29	1421.68	77.7
bsreoncp	809,340	10,759,259	6.77	15.92	2.4
catnaput	91,896	558,409	0.25	0.25	1.0
cndlckut	54,789	468,490	0.04	0.14	3.1
cpsbtfrc	360,069	4,057,273	1.98	2.87	1.4
dlhytrut	194,406	586,646	0.21	0.73	3.6
krhytrut	159,260	443,687	0.10	0.31	3.2
nphytrut	158,361	423,116	0.09	0.32	3.5
osothcut	424,072	1,442,526	0.87	2.83	3.3
otdlckut	55,241	504,224	0.05	0.15	3.2
hpdtcudm	265,616	3,355,299	1.95	5.27	2.7
tsrchcut	605,693	1,987,192	1.32	4.55	3.5

* Cold start is solution with no previous factorization info and results are sum of analyze + factorize + solve times

Copyright © 2024 by Honeywell International Inc. All rights reserved.

cuDSS PERFORMANCE SPEEDUP: HOT START*

Matrix	n	nnz(A)	CUDSS (secs)	DOTAXB (secs)	SpeedUp (DOTAXB/CUDSS)
lgcmpdis	1,136,993	76,789,656	5.19	1421.68	274.0
bsreoncp	809,340	10,759,259	0.57	15.92	28.0
catnaput	91,896	558,409	0.05	0.25	5.2
cndlckut	54,789	468,490	0.01	0.14	13.5
cpsbtfrc	360,069	4,057,273	0.16	2.87	18.0
dlhytrut	194,406	586,646	0.04	0.73	19.5
krhytrut	159,260	443,687	0.03	0.31	12.2
nphytrut	158,361	423,116	0.03	0.32	12.7
osothcut	424,072	1,442,526	0.12	2.83	23.1
otdlckut	55,241	504,224	0.01	0.15	14.1
hpdtcudm	265,616	3,355,299	0.16	5.27	33.5
tsrchcut	605,693	1,987,192	0.17	4.55	27.3

* Hot start is solution using previous factorization information and results are sum of refactorize + solve times

Copyright © 2024 by Honeywell International Inc. All rights reserved.

RESULTS SUMMARY

cuDSS is faster than DOTAXB in all cases with equivalent accuracy

Excluding the largest case, average cuDSS speedup

- 3X for cold start
- 19X for hot start

cuDSS performance speedup is greatest on the largest test case

- 78X for cold start
- 200X for hot start

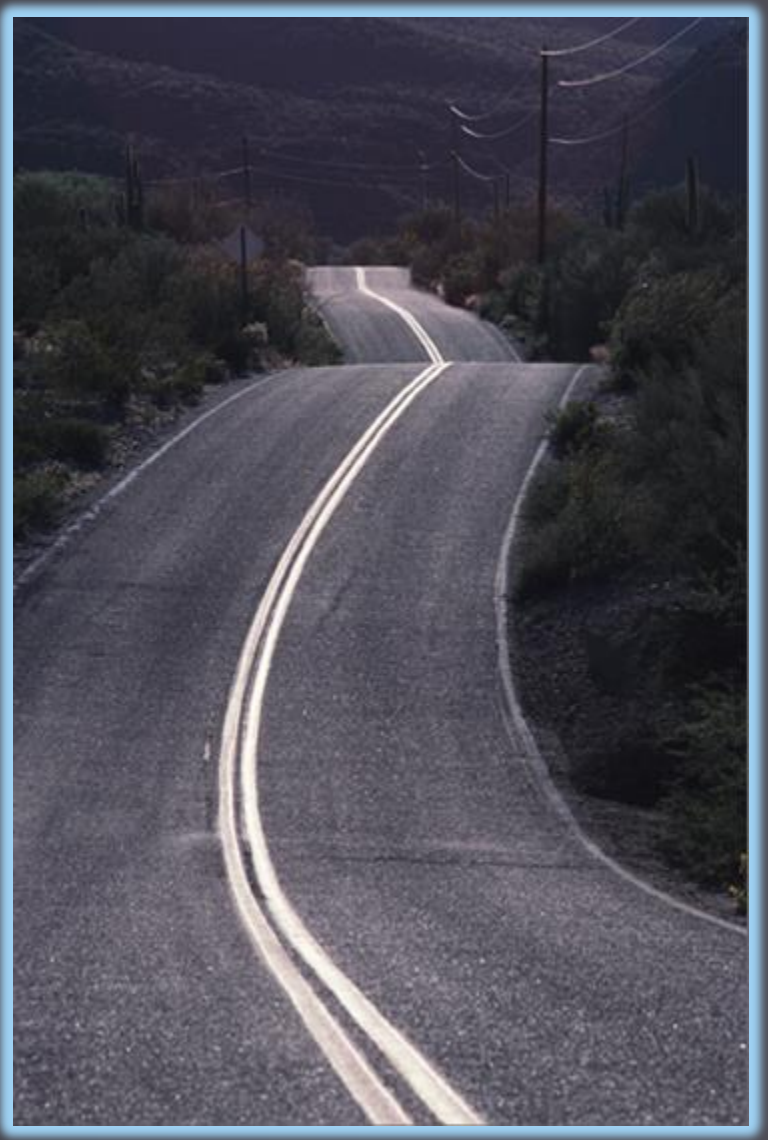
Copyright © 2024 by Honeywell International Inc. All rights reserved.

NEXT STEPS

Complete productization of cuDSS as a linear solver option within nonlinear equation solving in UniSim Design

Optimize cuDSS configuration for the process simulation domain

Assess improvements with different NVIDIA GPUs and new and emerging NVIDIA hardware



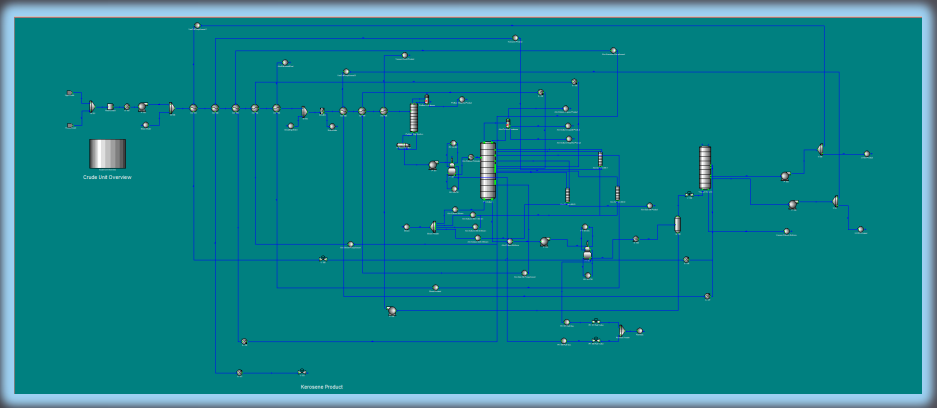
Copyright © 2024 by Honeywell International Inc. All rights reserved.

CONCLUSIONS

Modern simultaneous approaches to the solution of process simulation problems can exploit multicore CPU and GPU architectures through the sparse linear system solvers

NVIDIA’s cuDSS direct sparse linear system solver enables the use of GPUs to accelerate solution performance

The combination of simultaneous solution approaches and GPU-accelerated solvers can bring large performance improvements for simulation and optimization use cases in the process industries



Copyright © 2024 by Honeywell International Inc. All rights reserved.

QUESTIONS?

https://developer.nvidia.com/cudss

Copyright © 2024 by Honeywell International Inc. All rights reserved.