Real-time stochastic optimization of complex energy systems

Cosmin G. Petra — Argonne National Laboratory — petra@mcs.anl.gov

Olaf Schenk — Univ. della Svizzera italiana

Mihai Anitescu — Argonne National Laboratory

Stochastic optimization and the power grid

- Next generation energy systems assume a large penetration of renewable energy. Federal government goal is to produce 20% of the energy from renewable by 2030.
- Renewable energy is associated with highly variable weather conditions. The optimal electricity *dispatch decisions* must account for the *uncertainty* in the ambient factors.

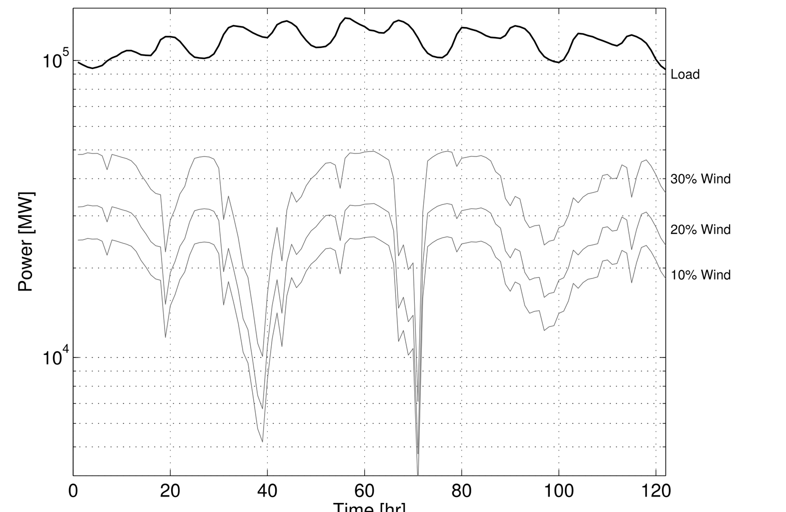


Figure 1: Drops in wind power more drastic than drops in demand.

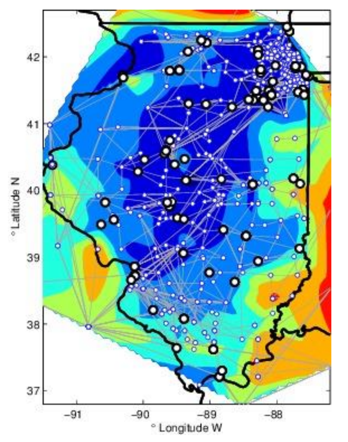


Figure 2: Illinois power grid

- Unit commitment (UC): optimal on/off schedule of thermal (coal, natural gas, nuclear) generators. Formulated as an MILP and solved hourly.
- Economic dispatch (ED): finer adjustments to generation levels and determine market prices. A continuous LP solved every 5-10 min.
- Current practice: deterministic. Uncertainty mitigated by having large back-up generation units (natural gas). It is not cost effective, increases the emissions and requires significant investments.

- **Challenge**: Integrate energy produced by highly variable renewable sources into these control systems.
- Consider a range of scenarios to find the optimal plant commitments. This results in a stochastic optimization problem with billions of variables.

PIPS - Parallel solver for stochastic optimization

Two-stage stochastic quadratic program (extensive form)

$$\min_{x,y_1,\dots,y_N} \frac{1}{2}x^TQx + c^Tx + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{2}y_i^TQ_iy_i + q_i^Ty_i\right)$$
$$\text{s.t. } Ax = b,$$
$$T_ix + W_iy_i = h_i,\ i = 1,\dots,N$$
$$x, y_1, \dots, y_N \geq 0.$$

N is the number of scenarios, can be as large as O(10,000)

Interior-point methods lead to block arrow-shaped linear systems

$$\begin{bmatrix} K_1 & & & B_1 \\ & \ddots & & \vdots \\ & & K_N & B_N \\ B_1^T & \dots & B_N^T & K_0 \end{bmatrix} \begin{bmatrix} \Delta z_1 \\ \vdots \\ \Delta z_N \\ \Delta z_0 \end{bmatrix} = \begin{bmatrix} r_1 \\ \vdots \\ r_N \\ r_0 \end{bmatrix}$$

$$K_0 = \begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \quad K_i = \begin{bmatrix} \bar{Q}_i & W_i^T \\ W_i & 0 \end{bmatrix}$$
$$B_i = \begin{bmatrix} 0 & 0 \\ T_i & 0 \end{bmatrix} \quad \Delta z_0 = \begin{bmatrix} x \\ \pi \end{bmatrix} \quad \Delta z_i = \begin{bmatrix} y_i \\ \lambda_i \end{bmatrix}$$

π, λ_i are eq. constr. multipliers

Algorithm 1. - Schur complement-based decomposition of the linear algebra

1. Calculate $B_i^TK_i^{-1}B_i,\ i = 1,\dots,N$ ("Compute S.C.") — ("embarrassingly" parallel)
2. Form $C := K_0 - \sum_{i=1}^N B_i^TK_i^{-1}B_i$ ("Form S.C.") — (communication - AllReduce)
3. Factorize $C = L_0D_0L_0^T$ ("Factor S.C.") — (computation bottleneck)
4. Solve $\Delta z_0 = C^{-1}(r_0 - \sum_{i=1}^N B_i^TK_i^{-1}r_i)$ — (comp. bottleneck, comm.)
5. Solve $\Delta z_i = K_i^{-1}(B_i\Delta z_0 - r_i),\ i = 1,\dots,N$ — ("embarrassingly" parallel)

The computational bottleneck can be reduced by using the stochastic preconditioner (**P.**, Anitescu 2012) or by distributing the dense Schur complement (Lubin, **P.** 2012). 90%+ strong scaling efficiency on 128K BG/P cores.

New algorithmic approach for the Schur complement computation

- Time-to-solution is very important on the operational side (UC needs to be solved every 1h).
- Classically, $B_i^TK_i^{-1}B_i$ is computed using repeated triangular solves with the factors of K_i
  - poor performance due to a memory bandwidth wall;
  - sparsity of data is not fully exploited.
- New **incomplete augmented factorization** algorithm
  - compute $B_i^TK_i^{-1}B_i$ by partially factorizing the augmented matrix $\begin{bmatrix} K_i & B_i^T \\ B_i & 0 \end{bmatrix}$
  - based on a supernodal factorization (in PARDISO); achieves good multi-threading performance and fully exploits the sparsity.
- Pivot perturbations are needed to maintain numerical stability (Bunch-Kauffman).
- **Preconditioned BiCGStab** is used to "absorb" the perturbations.

Algorithm 2. – Incomplete augmented factorization with BiCGStab error absorption

1. Calculate $B_i^T\tilde{K}_i^{-1}B_i$s, $i = 1,\dots,N$
2. Form $\tilde{C} := K_0 - \sum_{i=1}^N B_i^T\tilde{K}_i^{-1}B_i$
3. Factorize $\tilde{C} = L_0D_0L_0^T$
   (finished implicit factorization of $\tilde{K}$)
4. Solve for $K\Delta x = r$ using BiCGStab and $\tilde{K}^{-1}$ as preconditioner for K

   ⋮
   - Apply preconditioner $\tilde{K}^{-1}u$ by solving with $\tilde{K}$ as in Steps 4 and 5 of Alg. 1.
   - Compute the mat-vec Ku

   ⋮

   (end of BiCGStab iteration)

$$\tilde{K} := \begin{bmatrix} \tilde{K}_1 & & & B_1 \\ & \ddots & & \vdots \\ & & \tilde{K}_N & B_N \\ B_1^T & \dots & B_N^T & K_0 \end{bmatrix}$$

$$\tilde{K}_i = K_i + \Delta K_i$$

ΔK_i are the pivot perturbations

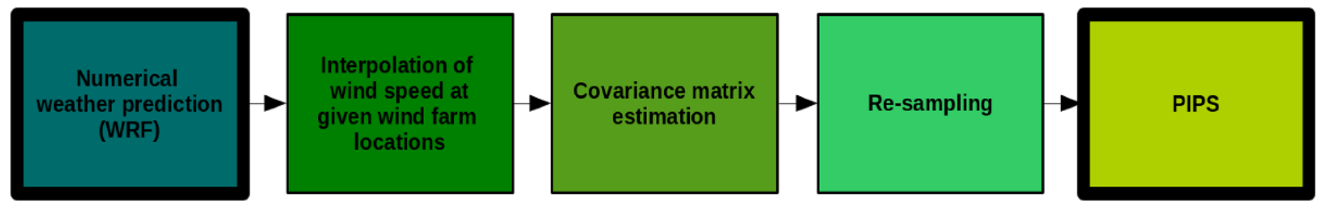
- Cost of one BiCGStab iteration is roughly equal to the cost of steps 4 and 5 in Algorithm 1. + the multiplication with K.
- Joint work with Olaf Schenk (USI, Lugano), Miles Lubin (MIT) and Klaus Gaertner (WIAS, Berlin)

Other recent computational/implementation advances

- Adoption of GPUs in the dense linear algebra computations (on XK7)
- Tune-up communication for Cray systems
- PIPS ported to IBM BG/Q, Cray XE6, Cray XK7, Cray XC30. Previously ran on IBM BG/P.

Integrative framework for the optimization of the power grid

Wind samples/ensembles are obtained using numerical weather prediction codes (WRF in particular). These simulations are computationally expensive, and only a small number samples can be afforded. We rely on resampling techniques. Joint work with Elias D. Nino (Virginia Tech).



Shrinkage estimators for the covariance matrix

$$\mathbf{S}_{\Theta} = \rho \cdot \mathbf{I} + (1-\rho)\cdot\mathbf{S}$$

Ledoit-Wolf $\rho = \min\left(\frac{\sum_{i=1}^n \mathbf{x}_i\cdot\mathbf{x}_i^t - \mathbf{S}}{n^2\left[\mathrm{tr}(\mathbf{S}^2) - \frac{\mathrm{tr}^2(\mathbf{S})}{p}\right]}, 1\right)$

Rao-Blackwell-Ledoit-Wolf $\rho = \min\left(\frac{\frac{n-2}{n}\cdot\mathrm{tr}\left(\mathbf{S}^2\right) + \mathrm{tr}^2\left(\mathbf{S}\right)}{(n+2)\cdot\left[\mathrm{tr}\left(\mathbf{S}^2\right) - \frac{\mathrm{tr}^2(\mathbf{S})}{p}\right]}, 1\right)$

$\mathbf{S} = \frac{1}{n}\cdot\left[\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})\cdot(\mathbf{x}_i - \bar{\mathbf{x}})^t\right] \in \Re^{p\times p}$ is the sample covariance matrix.

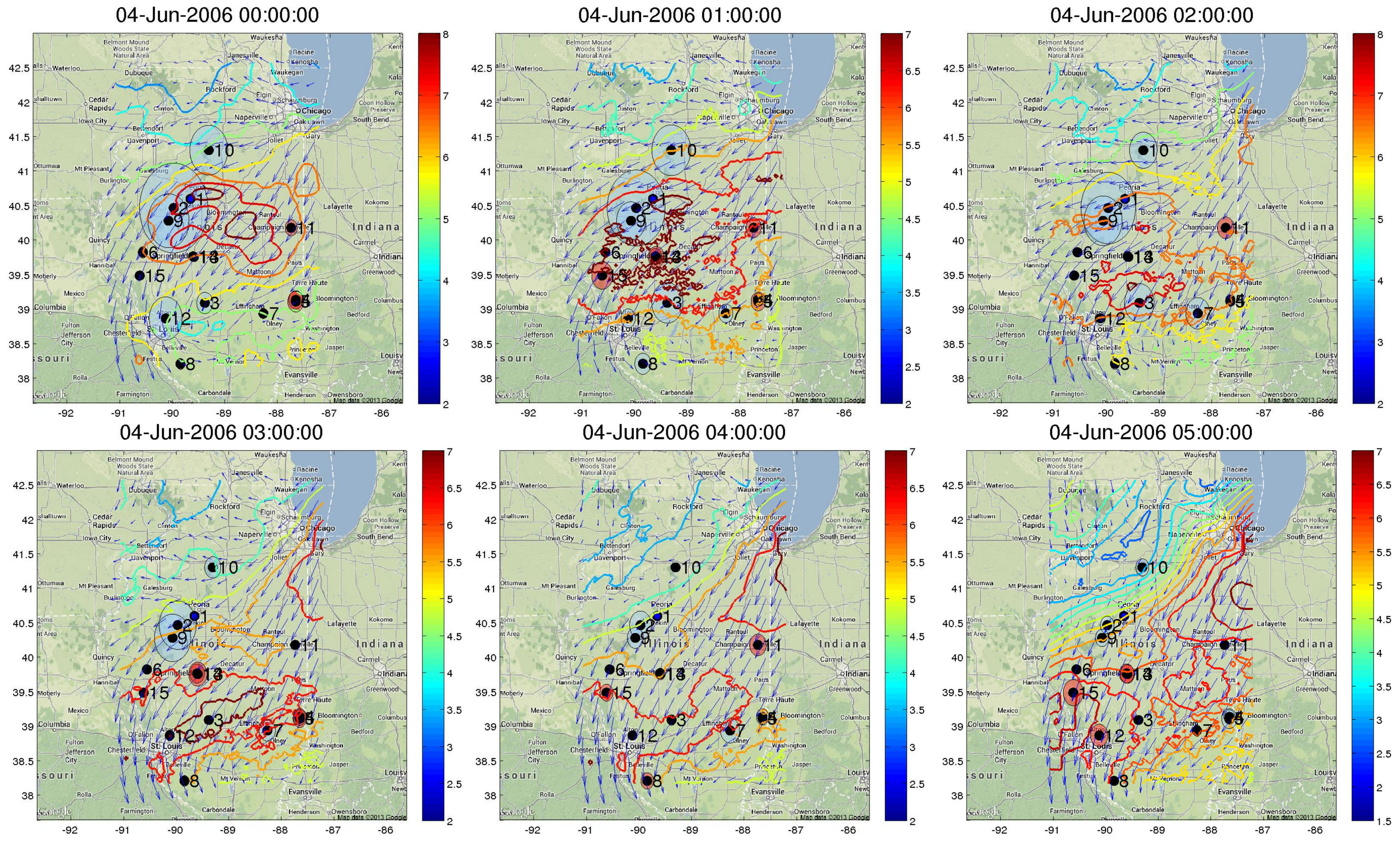


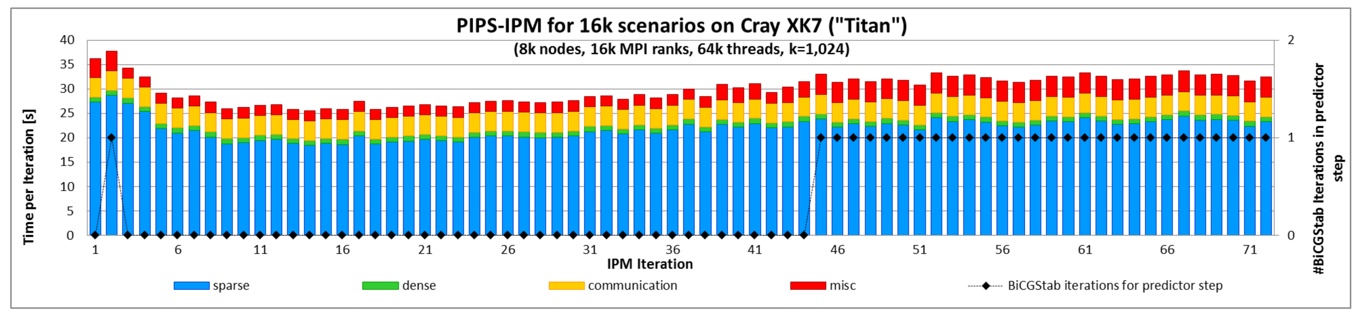
Figure 3: Wind correlations in time (reference is wind farm #1), wind directions and wind speed contours are shown.

Numerical experiments

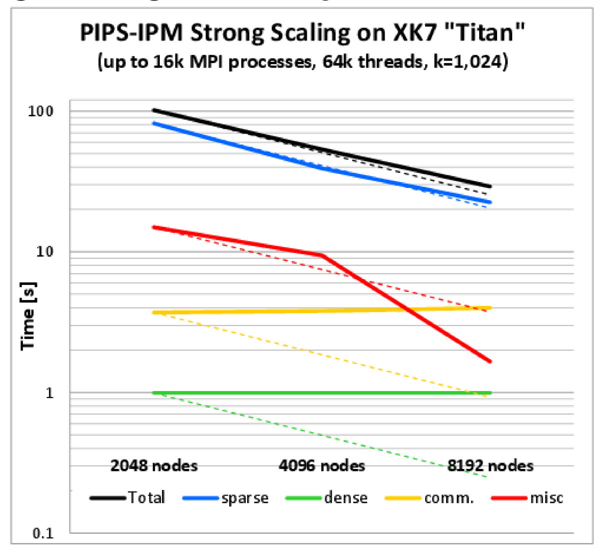
Test problems: UC with 24h horizon with the State of Illinois' network. Up to 2 billion decision variables and 2 billion constraints.

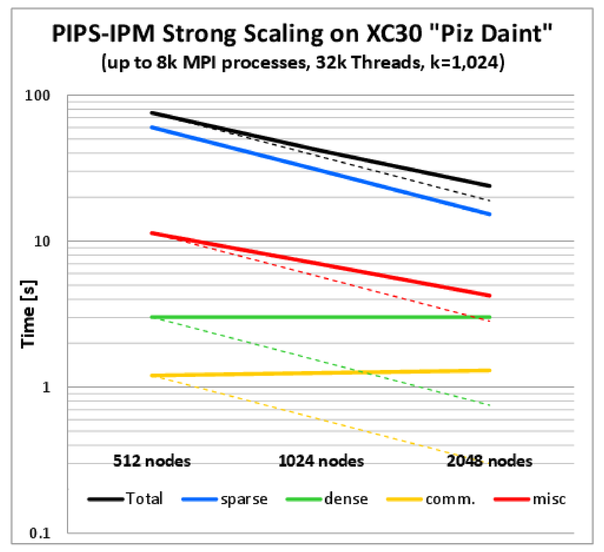
Architectures: "Titan" XK7 of Oak Ridge Natl. Lab, "Piz Daint" of Swiss Natl. Computing Centre.

Performance of the incomplete augmented factorization algorithm

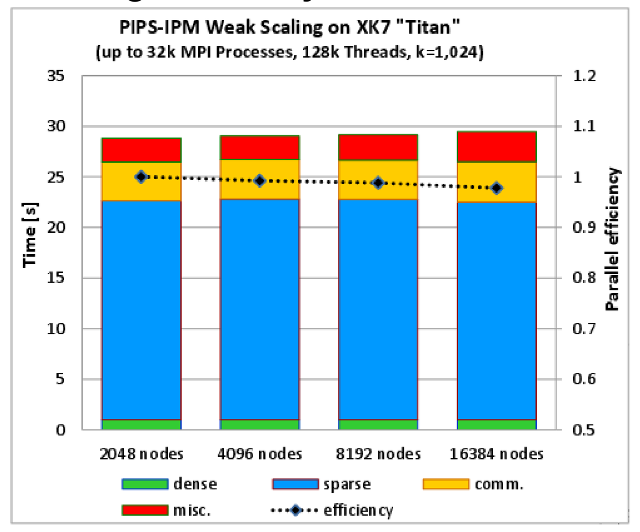


Strong scaling efficiency





Weak scaling efficiency



Intranode speedup and threads affinity

# of MPI ranks	(b) Number of threads per process on XC30: 1	1	2	4	8	16
	PARDISO	PARDISO-SC				
1	168.3	19.0	12.9	8.4	6.2	5.9
2	175.3	18.9	12.4	8.5	7.4	
4	194.6	19.5	13.8	12.1		
8	284.2	22.1	21.8			
16	281.6	38.5				

Conclusions and future work

- Large scale ED/UC relaxation can be solved in "real-time" on HPC platforms.
- Next: solve UC at the root with specialized cuts and feasibility pumps.
- Incorporate dynamical systems (AC power flow for example). Results in nonlinear nonconvex problems. New mathematics and algorithms are needed to ensure scalability.

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY