MultiFidelity UQ sampling for Stochastic Simulations

Gianluca Geraci, Laura P. Swiler and Bert J. Debusschere

16th U.S. National Congress on Computational Mechanics (USNCCM 16)
Virtual Event
July 26, 2021

Plan of the talk

- Background
- Network Modeling
- Stochastic solvers in Sampling Methods
- Numerical Examples
- Conclusions

Background

Uncertainty Quantification
DoE and DoD deployment activities

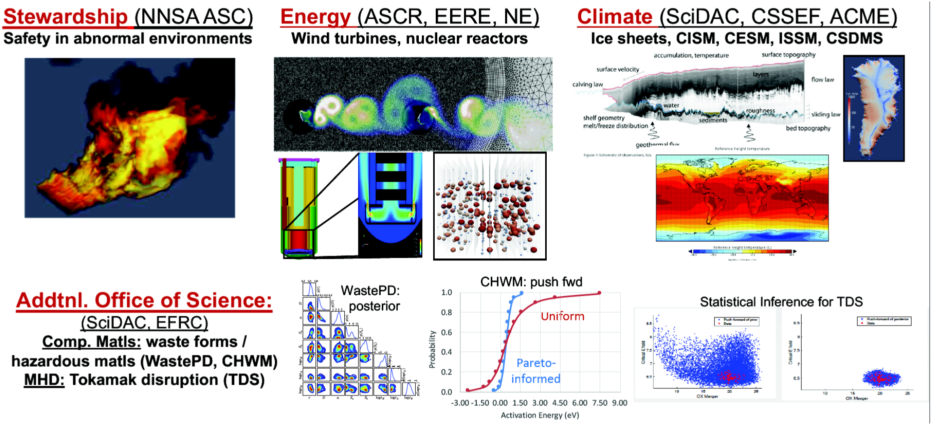


Figure courtesy of Mike Eldred

High-fidelity state-of-the-art modeling and simulations with HPC

- **Severe** simulations **budget constraints**
- **Significant dimensionality** driven by model complexity

Uncertainty Quantification
Forward Propagation – Why Sampling Methods?

UQ context at a glance:

- **Challenges:** High-dimensionality, non-linearity and possibly non-smooth responses
- **Opportunities:** Rich physics and several discretization levels/models available

Natural candidate:

- **Sampling**-based (MC-like) approaches because they are **non-intrusive**, **robust** and **flexible**...
- **Drawback:** Slow convergence $\mathcal{O}(N^{-1/2}) \rightarrow$ many realizations to build reliable statistics

Goal of MF UQ: **Reducing the computational cost** of obtaining MC reliable statistics

Pivotal idea:

- Simplified (**low-fidelity**) models are **inaccurate** but **cheap**
  - **low-variance** estimates
- **High-fidelity** models are **costly**, but **accurate**
  - **low-bias** estimates

Uncertainty Quantification
Forward Propagation – Why Sampling Methods?

UQ context at a glance:

- **Challenges:** High-dimensionality, non-linearity and possibly non-smooth responses
- **Opportunities:** Rich physics and several discretization levels/models available

Natural candidate:

- **Sampling**-based (MC-like) approaches because they are **non-intrusive**, **robust** and **flexible**...
- **Drawback:** Slow convergence $\mathcal{O}(N^{-1/2}) \rightarrow$ many realizations to build reliable statistics

Goal of MF UQ: **Reducing the computational cost** of obtaining MC reliable statistics

Pivotal idea:

- Simplified (**low-fidelity**) models are **inaccurate** but **cheap**
  - ► **low-variance** estimates
- **High-fidelity** models are **costly**, but **accurate**
  - ► **low-bias** estimates

Talk's Contribution: Addressing challenges (and opportunities) introduced by **Stochastic Solvers**

A comment on Stochastic solvers
Examples of Stochastic solvers at Sandia

Q: What do I mean in this presentation with **stochastic solver**?

A: For the sake of this presentation, I mean that in addition to the UQ parameters, the **solver has another source of variability** (that we cannot control)

Few **SNL-relevant examples**:

- **Turbulent flows/Combustion:** the stochasticity is introduced in the time-windowing used for statistics (we cannot integrate long enough)
- **Computer Networks/Cybersecurity:** virtualization of networks that runs real-time on specialized hardware, *i.e.* the status of the hardware produces background noise that cannot be controlled
- **Radiation transport:** MC transport solvers based on the propagation of a finite number of particle histories which need to be averaged to obtain the QoIs
- **GDSA:** stochasticity introduced in the (finite number of random realizations of the) subsurface modeling
- etc.

A comment on Stochastic solvers
Examples of Stochastic solvers at Sandia

Q: What do I mean in this presentation with **stochastic solver**?

A: For the sake of this presentation, I mean that in addition to the UQ parameters, the **solver has another source of variability** (that we cannot control)

Few **SNL-relevant examples**:

- **Turbulent flows/Combustion:** the stochasticity is introduced in the time-windowing used for statistics (we cannot integrate long enough)
- **Computer Networks/Cybersecurity:** virtualization of networks that runs real-time on specialized hardware, *i.e.* the status of the hardware produces background noise that cannot be controlled
- **Radiation transport:** MC transport solvers based on the propagation of a finite number of particle histories which need to be averaged to obtain the QoIs
- **GDSA:** stochasticity introduced in the (finite number of random realizations of the) subsurface modeling
- etc.

Talk's Focus: I'll demonstrate the use of MF UQ for stochastic solvers in the context of **Cybersecurity applications**

Network Modeling

Network models

Simulation vs Emulation: strengths and differences

Simulation: based on deep **understanding of the underlying processes**

- ✔ Fast to **develop**
- ✔ Runs **faster than real-time** since they control the clock
- ✔ Easy to **run in parallel**: neither time-dependent or reliant on virtualized hardware
- ✗ Unable to capture **emergent behaviors**

Emulation: able to **capture unknown or not well-understood behaviors**

- ✔ Runs the **real software** therefore closely resembles a physical testbed
- ✗ Requires **more hardware** and therefore the number of concurrent evaluations are limited

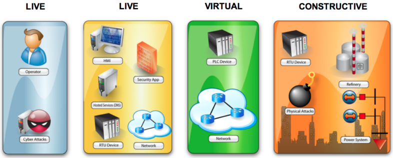


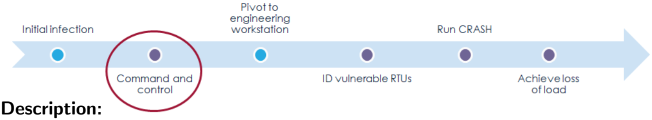
Figure courtesy of David Fritz, SAND2018-3927

Examples of Network modeling at Sandia (Courtesy of David Fritz, SAND2018-3927[1])

- **DevOps:** Ensure operation of new hardware, software, services in high-consequence environments. Detect malfunctions, misconfigurations and malicious consequences
- **Malware:** Understanding of malware through pseudo-in situ execution
- **ICS/SCADA:** Best countermeasures for my IT-connected ICS systems? Can we detect attacks? Can we assess resiliency of the IT-controls over the entire power grids?
- **Nuclear Weapons:** Assure Communication, Command and Control regardless of network state and threats?

[1] http://minimega.org/presentations/gt_2018.slide#7

Command & Control (C2) example
Problem Description

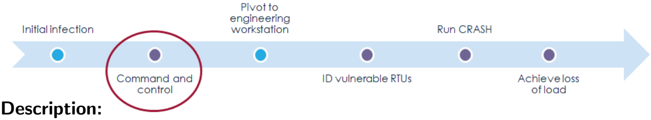


Problem Description:

- **Multi-stage attack** aiming at accessing a **power utility**'s corporate enterprise network
- **Attacker Goal:** pivoting on the industrial control system (ICS) to ultimately **cause load shed**
- **Attacker Perspective:** **Maintaining communication** between C2 host and C2 server
- **Defender Perspective:** Deploying an IDS to **identify/mitigate malicious traffic**

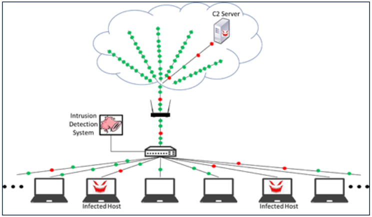
Command & Control (C2) example

Problem Description



Problem Description:

- **Multi-stage attack** aiming at accessing a **power utility**'s corporate enterprise network
- **Attacker Goal:** pivoting on the industrial control system (ICS) to ultimately **cause load shed**
- **Attacker Perspective:** **Maintaining communication** between C2 host and C2 server
- **Defender Perspective:** Deploying an IDS to **identify/mitigate malicious traffic**



Analysis Scenario:

- One or more hosts with the network have been infected
- **Benign** and **Malicious** traffic co-exist
- The **IDS performs packet inspection** and issues an alert if the content appears suspicious
- Large packet rates make malicious traffic difficult to detect
- **Emotet malware** (banking trojan from 2014) and the Snort IDS

Stochastic solvers in (MF) Sampling Methods

Few **definitions and notation**:

- ξ is the vector of **UQ parameters**
- η is a vector of inaccessible RV that notionally represents the **variability in the solver**
- Every time we run the solver, we get an **elementary realization** $f = f(\xi, \eta)$
- **Running for a fixed $\xi^{(i)}$** multiple times (*replicas*) generates $\left\{ f(\xi^{(i)}, \eta^{(j)}) \right\}_{j=1}^{N_\eta}$
- The QoI for UQ is obtained by **averaging f** (for a fixed ξ):

$$Q(\xi) = \mathbb{E}_\eta [f] \approx \frac{1}{N_\eta} \sum_{j=1}^{N_\eta} f(\xi^{(i)}, \eta^{(j)}) = \tilde{Q}(\xi)$$

Sampling UQ, *e.g.* mean estimator, is accomplished with **two nested sampling estimators**

$$\mathbb{E}[Q] \approx \frac{1}{N_\xi} \sum_{i=1}^{N_\xi} \tilde{Q}^{(i)} = \frac{1}{N_\xi} \sum_{i=1}^{N_\xi} \left[\frac{1}{N_\eta} \sum_{j=1}^{N_\eta} f(\xi^{(i)}, \eta^{(j)}) \right]$$

Q: How does the noise introduced in the finite averaging over *replicas* propagate in the estimator?

The MC estimator is still *unbiased*...

$$\hat{Q}^{MC} = \frac{1}{N_\xi}\sum_{i=1}^{N_\xi}\left(\frac{1}{N_\eta}\sum_{j=1}^{N_\eta} f(\xi^{(i)},\eta^{(j)})\right) = \frac{1}{N_\xi}\sum_{i=1}^{N_\xi}\tilde{Q}(\xi^{(i)}) \quad \longrightarrow \quad \mathbb{V}ar\left[\hat{Q}^{MC}\right] = \frac{\mathbb{V}ar\left[\tilde{Q}(\xi)\right]}{N_\xi}$$

Law of total variance

$$\mathbb{V}ar\left[\cdot\right] = \mathbb{V}ar\left[\mathbb{E}_\eta\left[\cdot\right]\right] + \mathbb{E}\left[\mathbb{V}ar_\eta\left[\cdot\right]\right],$$

It follows that

$$\begin{aligned}\mathbb{V}ar\left[\tilde{Q}(\xi;\eta)\right] &= \mathbb{V}ar\left[\mathbb{E}_\eta\left[\tilde{Q}(\xi;\eta)\right]\right] + \mathbb{E}\left[\mathbb{V}ar_\eta\left[\tilde{Q}(\xi;\eta)\right]\right] \\ &= \mathbb{V}ar\left[\mathbb{E}_\eta\left[f(\xi,\eta)\right]\right] + \frac{\mathbb{E}\left[\mathbb{V}ar_\eta\left[f(\xi,\eta)\right]\right]}{N_\eta} \\ &= \mathbb{V}ar\left[Q(\xi)\right] + \mathbb{E}\left[\frac{\sigma_\eta^2(\xi)}{N_\eta}\right].\end{aligned}$$

Finally,

$$\mathbb{V}ar\left[\hat{Q}^{MC}\right] = \frac{\mathbb{V}ar\left[\mathbf{Q}(\xi)\right] + \mathbb{E}\left[\frac{\sigma_\eta^2(\xi)}{\mathbf{N}_\eta}\right]}{N_\xi}$$

NOTES:

- The **true variance** is augmented by the (average) **noise** introduced by replicas
- The average **noise** is the average variance of the inner MC estimator

Let's now consider the **control variate** case[2]:

$$\hat{Q}^{CV} = \hat{Q}^{MC}_{HF} + \alpha \left(\hat{Q}^{MC}_{LF} - \mathbb{E}\left[Q^{LF} \right] \right),$$

for which we know a solution that minimizes the estimator cost (for a prescribed variance ε^2).

For the **control variate** case, we need to consider the following **properties**:

- All quantities depend on the **number of replicas**, *i.e.* N^{HF}_{η} and N^{LF}_{η}
- The **correlation** between quantities $\tilde{Q}^{HF}$ and $\tilde{Q}^{LF}$ **increases** by averaging more replicas
- The **computational cost** **increases** with the number of replicas

Q: Can we **optimize** the number of replicas in order to **maximize the efficiency** of the estimator?

[2] This is equivalent to MFMC or ACV with one low-fidelity model

Let's start from the **classical control variate solution** given the quantities $\tilde{Q}^{HF}$ and $\tilde{Q}^{LF}$:

$$\mathbb{V}ar\left[\hat{Q}^{CV}\right] = \frac{\mathbb{V}ar\left[\tilde{Q}^{HF}\right]}{N_\xi}\left(1 - \frac{\tilde{r}-1}{\tilde{r}}\tilde{\rho}^2\right), \quad \text{with} \quad \tilde{r} = \sqrt{\frac{\tilde{\mathcal{C}}_{HF}}{\tilde{\mathcal{C}}_{LF}}\frac{\tilde{\rho}^2}{1-\tilde{\rho}^2}},$$

where we use a **total number of low-fidelity simulations** equal to $\lceil \tilde{r} N_\xi \rceil$.

Stochastic Solvers and sampling methods
Extension to Multifidelity (2/3)

Let's start from the **classical control variate solution** given the quantities $\tilde{Q}^{HF}$ and $\tilde{Q}^{LF}$:

$$\mathbb{V}ar\left[\hat{Q}^{CV}\right] = \frac{\mathbb{V}ar\left[\tilde{Q}^{HF}\right]}{N_\xi}\left(1 - \frac{\tilde{r}-1}{\tilde{r}}\tilde{\rho}^2\right), \quad \text{with} \quad \tilde{r} = \sqrt{\frac{\tilde{\mathcal{C}}_{HF}}{\tilde{\mathcal{C}}_{LF}}\frac{\tilde{\rho}^2}{1-\tilde{\rho}^2}},$$

where we use a **total number of low-fidelity simulations** equal to $\lceil \tilde{r}N_\xi \rceil$.

Q: Can we write this solution such that we **separate the stochastic component** controlled by the number of replicas?

Stochastic Solvers and sampling methods
Extension to Multifidelity (2/3)

Let's start from the **classical control variate solution** given the quantities $\tilde{Q}^{HF}$ and $\tilde{Q}^{LF}$:

$$\mathbb{V}ar\left[\hat{Q}^{CV}\right] = \frac{\mathbb{V}ar\left[\tilde{Q}^{HF}\right]}{N_\xi}\left(1 - \frac{\tilde{r}-1}{\tilde{r}}\tilde{\rho}^2\right), \quad \text{with} \quad \tilde{r} = \sqrt{\frac{\tilde{\mathcal{C}}_{HF}}{\tilde{\mathcal{C}}_{LF}}\frac{\tilde{\rho}^2}{1-\tilde{\rho}^2}},$$

where we use a **total number of low-fidelity simulations** equal to $\lceil \tilde{r} N_\xi \rceil$.

Q: Can we write this solution such that we **separate the stochastic component** controlled by the number of replicas?

STEP 1: **Re-write the correlation** as

$$\tilde{\rho}^2 = \frac{\left(\mathrm{Cov}\left[\tilde{Q}^{HF}, \tilde{Q}^{LF}\right]\right)^2}{\mathbb{V}ar\left[\tilde{Q}^{HF}\right]\mathbb{V}ar\left[\tilde{Q}^{LF}\right]} \longrightarrow \boxed{\tilde{\rho}^2 = \frac{\rho^2}{1+\rho^2\tilde{\tau}}},$$

$$\text{where} \quad \tilde{\tau} = \frac{\mathbb{V}ar\left[Q^{LF}\right]\frac{\mathbb{E}\left[\sigma^2_{\eta,\mathrm{HF}}\right]}{N_\eta^{\mathrm{HF}}} + \mathbb{V}ar\left[Q^{HF}\right]\frac{\mathbb{E}\left[\sigma^2_{\eta,\mathrm{LF}}\right]}{N_\eta^{\mathrm{LF}}} + \frac{\mathbb{E}\left[\sigma^2_{\eta,\mathrm{HF}}\right]\mathbb{E}\left[\sigma^2_{\eta,\mathrm{LF}}\right]}{N_\eta^{\mathrm{HF}}N_\eta^{\mathrm{LF}}}}{\left(\mathrm{Cov}\left[\tilde{Q}^{HF}, \tilde{Q}^{LF}\right]\right)^2}$$

NOTES:

- $\mathrm{Cov}\left[\tilde{Q}^{HF}, \tilde{Q}^{LF}\right] = \mathrm{Cov}\left[Q^{HF}, Q^{LF}\right]$
- $\mathbb{V}ar\left[\tilde{Q}^{HF}\right] = \mathbb{V}ar\left[Q^{HF}\right] + \mathbb{E}\left[\frac{\sigma_\eta^{2,HF}}{N_\eta}\right]$ (same for the low-fidelity)

STEP 2: **Re-write the cost** as a function of an elementary realization

$$\tilde{\mathcal{C}}_{\mathrm{HF}} = N_\eta^{\mathrm{HF}} \mathcal{C}^{\mathrm{HF}}$$
$$\tilde{\mathcal{C}}_{\mathrm{LF}} = N_\eta^{\mathrm{LF}} \mathcal{C}^{\mathrm{LF}}$$

STEP 3: Finally **write the sample allocation** (for a prescribed variance ε^2)

$$\tilde{r}^\star = \sqrt{\frac{\mathbf{1}-\boldsymbol{\rho}^{\mathbf{2}}}{\mathbf{1}-\boldsymbol{\rho}^{\mathbf{2}}+\boldsymbol{\rho}^{\mathbf{2}}\tilde{\tau}} \frac{\mathbf{N}_\eta^{\mathrm{HF}}}{\mathbf{N}_\eta^{\mathrm{LF}}}} \sqrt{\frac{\rho^2}{1-\rho^2}\frac{\mathcal{C}^{\mathrm{HF}}}{\mathcal{C}^{\mathrm{LF}}}} = \tilde{\mathbf{R}} r^\star$$

$$\tilde{\Lambda} = 1 - \frac{\tilde{R}r^\star - 1}{\tilde{R}r^\star} \frac{\rho^2}{1+\rho^2\tilde{\tau}}$$

$$N_\xi^\star = \frac{\mathbb{V}ar\left[Q^{HF}\right] + \frac{1}{N_\eta^{\mathrm{HF}}}\mathbb{E}\left[\sigma^2_{\eta,\mathrm{HF}}\right]}{\varepsilon^2}\tilde{\Lambda}$$

$$\mathcal{C}_{tot} = N_\xi \tilde{\mathcal{C}}_{\mathrm{HF}} + \tilde{r} N_\xi \tilde{\mathcal{C}}_{\mathrm{LF}} = N_\xi \mathcal{C}^{\mathrm{HF}} \left(N_\eta^{\mathrm{HF}} + \tilde{R} r \frac{\mathcal{C}^{\mathrm{LF}}}{\mathcal{C}^{\mathrm{HF}}} N_\eta^{\mathrm{LF}} \right)$$

STEP 2: **Re-write the cost** as a function of an elementary realization

$$\tilde{\mathcal{C}}_{\mathrm{HF}} = N_\eta^{\mathrm{HF}} \mathcal{C}^{\mathrm{HF}}$$
$$\tilde{\mathcal{C}}_{\mathrm{LF}} = N_\eta^{\mathrm{LF}} \mathcal{C}^{\mathrm{LF}}$$

STEP 3: Finally **write the sample allocation** (for a prescribed variance ε^2)

$$\tilde{r}^\star = \sqrt{\frac{1-\rho^2}{1-\rho^2+\rho^2\tilde{\tau}} \frac{N_\eta^{\mathrm{HF}}}{N_\eta^{\mathrm{LF}}}} \sqrt{\frac{\rho^2}{1-\rho^2} \frac{\mathcal{C}^{\mathrm{HF}}}{\mathcal{C}^{\mathrm{LF}}}} = \tilde{R}r^\star \quad \leftarrow \textbf{LF oversampling}$$

$$\tilde{\Lambda} = 1 - \frac{\tilde{R}r^\star - 1}{\tilde{R}r^\star} \frac{\rho^2}{1+\rho^2\tilde{\tau}} \quad \leftarrow \textbf{variance reduction}$$

$$N_\xi^\star = \frac{\mathbb{V}ar\left[Q^{HF}\right] + \frac{1}{N_\eta^{\mathrm{HF}}}\mathbb{E}\left[\sigma^2_{\eta,\mathrm{HF}}\right]}{\varepsilon^2} \tilde{\Lambda} \quad \leftarrow \textbf{HF samples}$$

$$\mathcal{C}_{tot} = N_\xi \tilde{\mathcal{C}}_{\mathrm{HF}} + \tilde{r} N_\xi \tilde{\mathcal{C}}_{\mathrm{LF}} = N_\xi \mathcal{C}^{\mathrm{HF}} \left(N_\eta^{\mathrm{HF}} + Rr \frac{\mathcal{C}^{\mathrm{LF}}}{\mathcal{C}^{\mathrm{HF}}} N_\eta^{\mathrm{LF}} \right) \quad \leftarrow \textbf{Total cost}$$

NOTE: All quantities denoted with $\tilde{\cdot}$ depend on the number of replicas

MF for stochastic solvers

Can we explore the efficiency of the estimator?

The total cost of the MF estimator (for reaching a prescribed variance ε^2) is

$$\mathcal{C}_{tot}^{MF} = \frac{\mathbb{V}ar\left[f^{\mathrm{HF}}\right] + \frac{1}{\mathbf{N}_\eta^{\mathrm{HF}}}\mathbb{E}\left[\sigma^2_{\eta,\mathrm{HF}}\right]}{\varepsilon^2}\mathcal{C}^{\mathrm{HF}}\left(1 - \frac{\tilde{\mathbf{R}}r - 1}{\tilde{\mathbf{R}}r}\frac{\rho^2}{1+\rho^2\tilde{\tau}}\right)\left(\mathbf{N}_\eta^{\mathrm{HF}} + \tilde{\mathbf{R}}r\frac{\mathcal{C}^{\mathrm{LF}}}{\mathcal{C}^{\mathrm{HF}}}\mathbf{N}_\eta^{\mathrm{LF}}\right)$$

Q: How costly would it be to obtain the same accuracy with MC?

$$\mathcal{C}_{MC} = \frac{\mathbb{V}ar\left[f^{\mathrm{HF}}\right] + \frac{1}{N_\eta^{\mathrm{HF}}}\mathbb{E}\left[\sigma^2_{\eta,\mathrm{HF}}\right]}{\varepsilon^2}\tilde{\mathcal{C}}_{\mathrm{HF}} = \frac{\mathbb{V}ar\left[f^{\mathrm{HF}}\right] \frac{1}{N_\eta^{\mathrm{HF}}}\mathbb{E}\left[\sigma^2_{\eta,\mathrm{HF}}\right]}{\varepsilon^2}\mathcal{C}^{\mathrm{HF}}N_\eta^{\mathrm{HF}}$$

Cost ratio:

$$\Theta = \frac{\mathcal{C}_{tot}^{MF}}{\mathcal{C}_{MC}} = \left(1 - \frac{\tilde{R}r - 1}{\tilde{R}r}\frac{\rho^2}{1+\rho^2\tilde{\tau}}\right)\left(1 + \tilde{R}r\frac{\mathcal{C}^{\mathrm{LF}}}{\mathcal{C}^{\mathrm{HF}}}\frac{N_\eta^{\mathrm{LF}}}{N_\eta^{\mathrm{HF}}}\right)$$

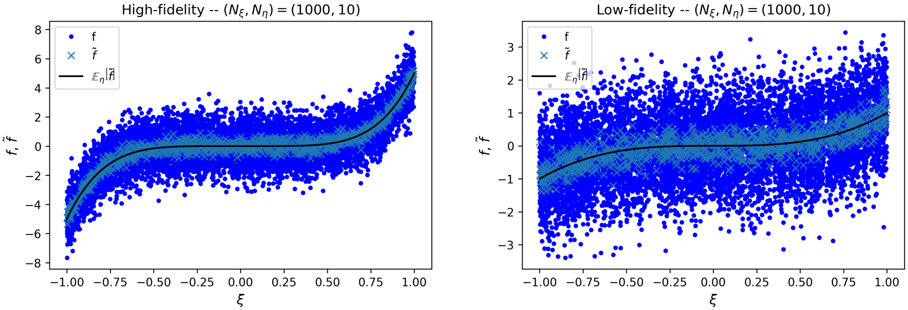
For all these results, in the limit of no noise and $N_\eta^{\mathrm{HF}} = N_\eta^{\mathrm{LF}} = 1$ the original allocation problem is recovered (HINT: $\tau = 0$ and $\tilde{\rho}^2 = \rho^2$...)

Numerical Examples

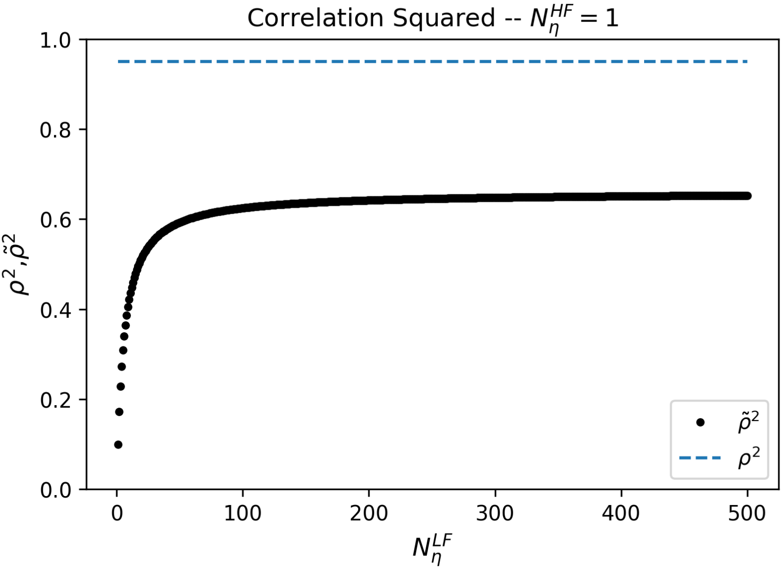
Analytical Verification

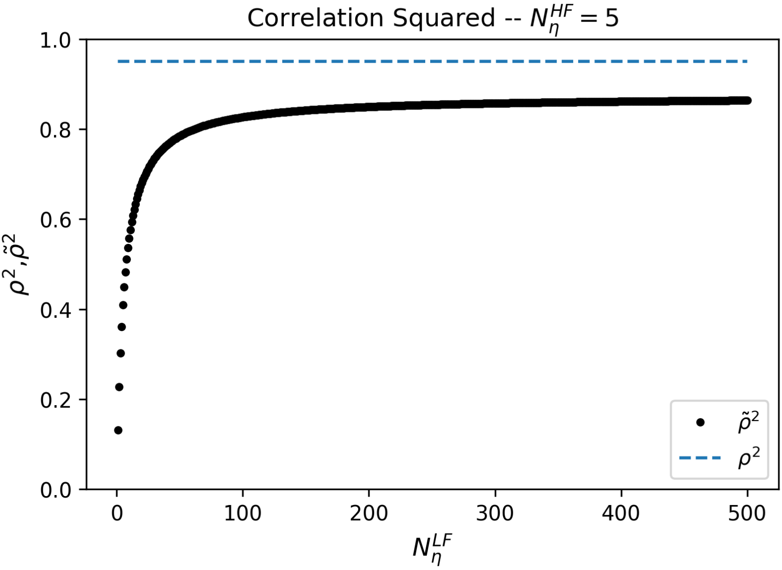
Verification Test Case
Problem Definition

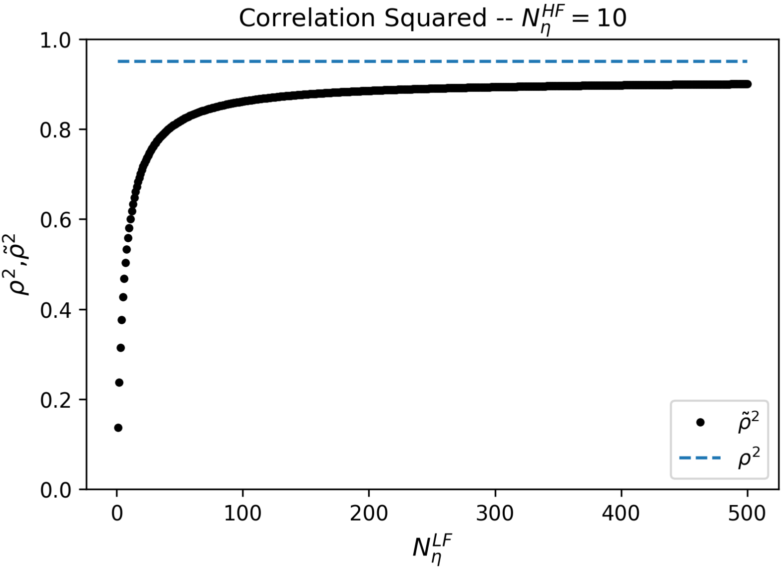
- Stochastic Parameter: $\xi \sim \mathcal{U}(-1, 1)$
- $f_{\text{HF}} = 5\xi^5 + \eta_{\text{HF}}$ where $\eta_{\text{HF}} \sim \mathcal{N}(0, \sigma_{\eta,\text{HF}} = 1)$
- $f_{\text{LF}} = \xi^3 + \eta_{\text{LF}}$ where $\eta_{\text{LF}} \sim \mathcal{N}(0, \sigma_{\eta,\text{LF}} = 0.9)$

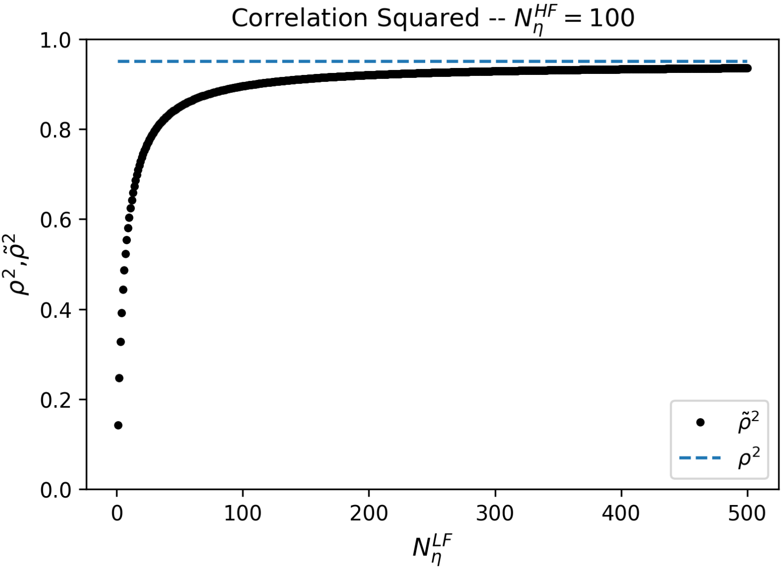


Q: Can we explore the **effect of the number of replicas on the correlation**?
A: The polluted correlation $\tilde{\rho}^2$ is expected to approach ρ^2 for $(N_\eta^{HF}, N_\eta^{LF}) \to \infty$


Correlation Squared -- $N_\eta^{HF} = 1$
1.0
0.8
0.6
0.4
0.2
0.0
$\rho^2, \tilde{\rho}^2$
0
100
200
300
400
500
N_η^{LF}
$\tilde{\rho}^2$
ρ^2


Correlation Squared -- $N_\eta^{HF} = 5$
1.0
0.8
0.6
0.4
0.2
0.0
$\rho^2, \tilde{\rho}^2$
0
100
200
300
400
500
N_η^{LF}
$\tilde{\rho}^2$
ρ^2

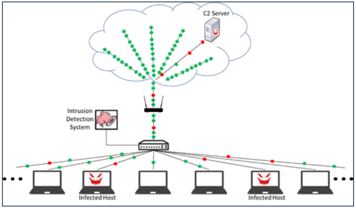

Correlation Squared -- $N_\eta^{HF} = 10$
1.0
0.8
0.6
0.4
0.2
0.0
$\rho^2, \tilde{\rho}^2$
0
100
200
300
400
500
N_η^{LF}
$\tilde{\rho}^2$
ρ^2



Correlation Squared -- $N_\eta^{HF} = 100$
1.0
0.8
0.6
0.4
0.2
0.0
$\rho^2, \tilde{\rho}^2$
0
100
200
300
400
500
N_η^{LF}
$\tilde{\rho}^2$
ρ^2

Cybersecurity – Command and Control (C2)

Command & Control (C2) example
Multifidelity UQ Analysis Scenario



Computer Models:

- **High-Fidelity** is a performed via a **Emulation-Based model**:
  - Set of Virtual Machines (VMs) running **full operating systems on virtualized hardware**
  - `minimega` (SNL) tool for launching and managing VMs
  - SCORCH (SNL) automated scenario orchestration framework
  - Background traffic is present in the system (it introduces stochastic behavior)
  - **Runtime:** 162 s
- **Low-Fidelity** is a performed via a **math model**:
  - **Probabilistic and discrete-time** representation for both traffic and IDS
  - **Runtime:** 0.001 s
  - **Cost ratio** $\mathcal{C}^{HF}/\mathcal{C}^{LF}$ is very high (162×10^3)

Parameters Varied in Experiment	Units	Value	Distribution
Aggregate Benign traffic rate in packets/sec	Packets per sec	100-3000	Continuous log-uniform
Fraction of benign packets with emotet signatures.	No Units	1e-5 to 8e-4	Log-uniform
Aggregate Malware traffic rate in packets/sec	Packets per sec	10-20	Uniform
Fraction of malware packets with emotet signatures	No units	0.01-0.025	Uniform
RAM assigned to the 1 CPU running SNORT	Mbytes	128, 256, 512, 1024	Discrete with equal probability

Command & Control example
Numerical Approach

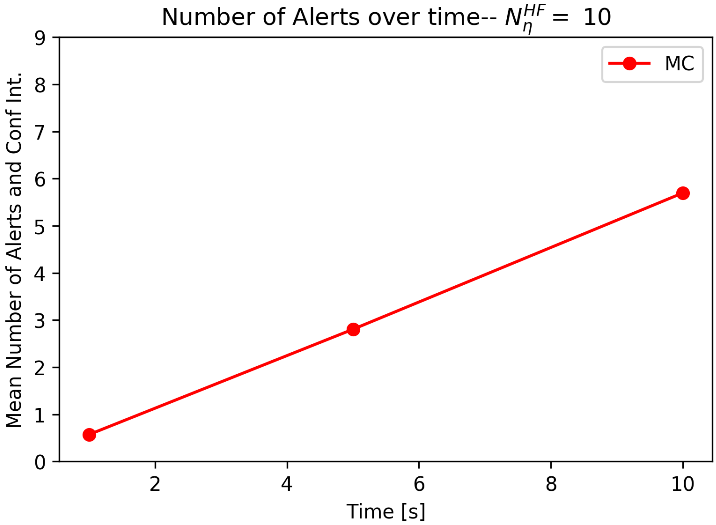
Numerical Study:

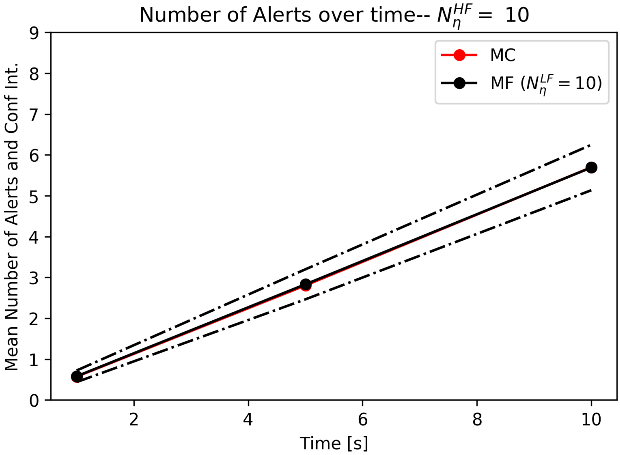
- We have a pilot set of $(N_\xi, N_\eta^{HF}) = (40, 10)$ for both model
  - From pilot samples we can estimate $\rho^2 = \rho^2(\tilde{\tau}, \tilde{\rho})$
  - $\mathbb{V}ar\left[Q^{HF}\right] = \mathbb{V}ar\left[Q^{HF}\right]\left(\mathbb{V}ar\left[\tilde{Q}^{HF}\right], N_\eta^{HF}\right)$ and $\mathbb{V}ar\left[Q^{LF}\right] = \mathbb{V}ar\left[Q^{LF}\right]\left(\mathbb{V}ar\left[\tilde{Q}^{LF}\right], N_\eta^{LF}\right)$
- We want to **build the most optimal MF estimator given the HF runs**, *i.e.* we need to optimize the total number of LF simulations (and the number of replicas)

Optimization Solutions:

1. MF estimator given $N_\eta^{LF} = 10$: the total number of LF runs is $\lceil \tilde{R}(10, 10)\, r^\star \rceil \times 10$
2. MF estimator with optimal N_η^{LF}:
$$\underset{N_\eta^{LF}}{\operatorname{argmin}}\, \Theta$$
The total number of LF runs is $\lceil \tilde{R}(10, N_\eta^{LF,\star})\, r^\star \rceil \times N_\eta^{LF,\star}$
3. MF estimator with $N_\eta^{LF} = 10$ but total cost equal to **[2]**: $\lceil \tilde{R}(10, N_\eta^{LF,\star})\, \mathbf{r} \rceil \times N_\eta^{LF,\star}$
4. MC with equivalent cost (of **[2]** and **[3]**)

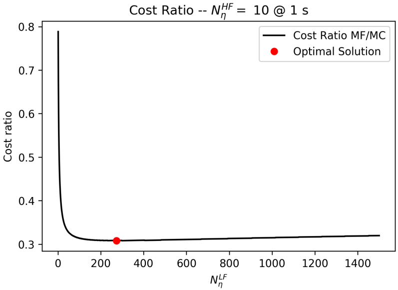
We consider three temporal locations, *i.e.* 1, 5, 10 s → **select the most restrictive condition** (highest number of LF runs)

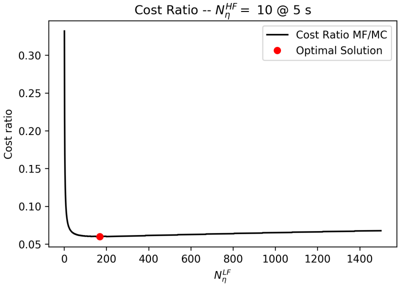

Number of Alerts over time-- $N_{\eta}^{HF} =$ 10
MC
Mean Number of Alerts and Conf Int.
Time [s]
0
1
2
3
4
5
6
7
8
9
2
4
6
8
10

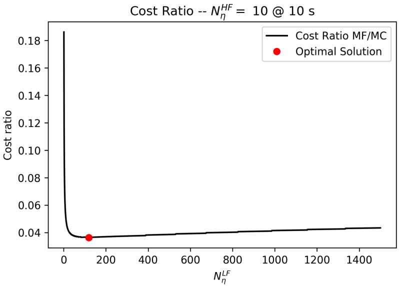


Estimator	N_ξ	$\tilde{r}$	N_η^{LF}	$\tilde{\Lambda}$ $t = 1$s	$t = 5$s	$t = 10$s	$N_\xi^{HF,eq}$
MF	40	1762.34	10	0.425	0.091	0.050	41

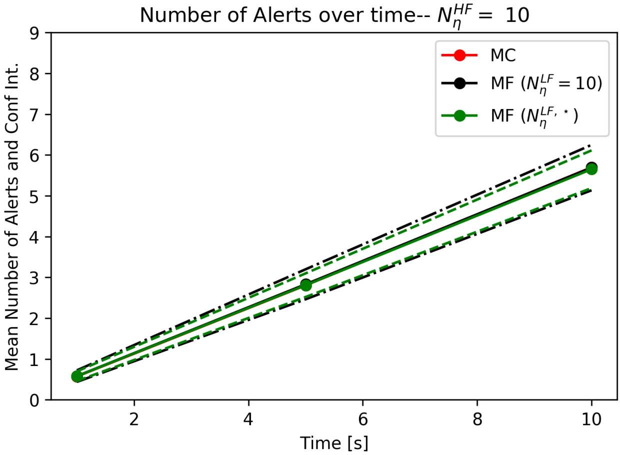
Command & Control example
Optimizing Number of Replicas







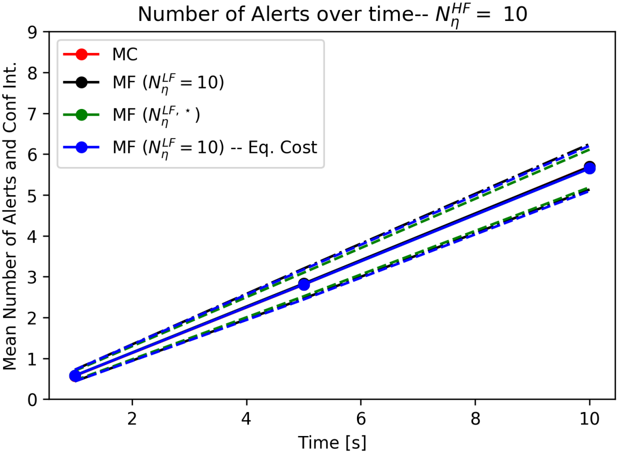
Solving N_η^{LF} for the minimum cost ratio, i.e. maximum MF efficiency



Estimator	N_ξ	$\tilde{r}$	N_η^{LF}	$\tilde{\Lambda}$ $t = 1$s	$t = 5$s	$t = 10$s	$N_\xi^{HF,eq}$
MF	40	1762.34	10	0.425	0.091	0.050	41
MF ($N_\eta^{LF,\star}$)	40	421	274	0.297	0.056	0.034	43

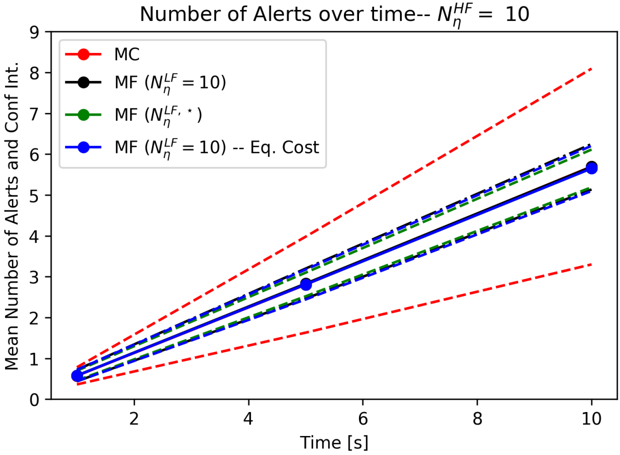
Command & Control example

Comparing Optimization Strategies – 99.7% Confidence Interval



Estimator	N_ξ	$\tilde{r}$	N_η^{LF}	$\tilde{\Lambda}$ $t = 1$s	$t = 5$s	$t = 10$s	$N_\xi^{HF,eq}$
MF	40	1762.34	10	0.425	0.091	0.050	41
MF ($N_\eta^{LF,\star}$)	40	421	274	0.297	0.056	0.034	43
MF ($\mathcal{C}_{eq}$)	40	11536.8	10	0.424	0.090	0.049	43

Command & Control example
Comparing Optimization Strategies – 99.7% Confidence Interval

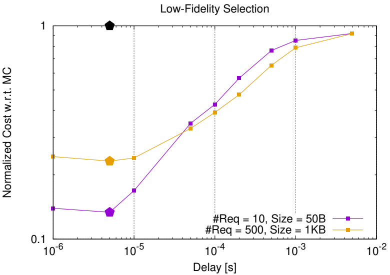


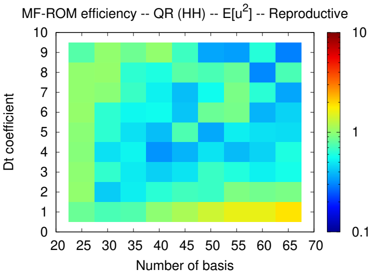
Estimator	N_ξ	$\tilde{r}$	N_η^{LF}	$\tilde{\Lambda}$ $t = 1$s	$t = 5$s	$t = 10$s	$N_\xi^{HF,eq}$
MF	40	1762.34	10	0.425	0.091	0.050	41
MF ($N_\eta^{LF,\star}$)	40	421	274	0.297	0.056	0.034	43
MF ($\mathcal{C}_{eq}$)	40	11536.8	10	0.424	0.090	0.049	43
MC	43	-	-	-	-	-	43

Conclusions

Is something similar arising in other contexts?

- This problem can be interpreted as an instance of a **Model Tuning** problem for MF
- **Model Tuning** in this context means that you could have additional parameters non-shared among models that you could consider hyper-parameters for tuning by increasing the correlation among models, and hopefully the MF estimator efficiency
- For instance, you could select the best spatial resolution of a LF model, the 'optimal' RANS coefficients in a LES-RANS MF problem, etc.
- In a **Model Tuning** exercise you can only optimize LF model parameters





Summary:

- **Stochastic Solvers** are widely used for several computational applications
- UQ with Sampling-based methods needs to **incorporate the noise** introduced by these solvers
- We demonstrated that the **number of replicas can be optimized** for maximizing the MF estimator efficiency

Work-in-Progress:

- Maximizing the MF efficiency is a Model Tuning exercise (more on this in Mike Eldred's talk)
- Extension to ACV (or similar approaches) is possible (and possibly more interesting)

(Incomplete) list of references:

CV Pasupathy, R., Taaffe, M., Schmeiser, B. W. & Wang, W., Control-variate estimation using estimated control means. *IIE Transactions*, **44**(5), 381–385, 2012

MFMC Ng, L.W.T. & Willcox, K. Multifidelity Approaches for Optimization Under Uncertainty. *Int. J. Numer. Meth. Engng* 100, no. 10, pp. 746772, 2014.

MFMC Peherstorfer, B., Willcox, K. & Gunzburger, M., Optimal Model Management for Multifidelity Monte Carlo Estimation. *SIAM J. Sci. Comput*. 38(5), A3163A3194.

ACV A.A. Gorodetsky, G. Geraci, M.S. Eldred & J.D. Jakeman, A Generalized Framework for Approximate Control Variates. *Journal of Computational Physics*, 2020.

CompNetw G. Geraci, J. Crussell, L.P. Swiler & B.J. Debusschere, Exploration of multifidelity UQ sampling strategies for computer network applications. *International Journal for Uncertainty Quantification*, Vol. 11(1), 2021.

MF-ROM P. Blonigan, G. Geraci, F. Rizzi, M.S. Eldred, Towards an integrated and efficient framework for leveraging reduced order models for multifidelity uncertainty quantification. *AIAA SciTech 2020*, 2020.

THANKS!

Acknowledgements

- Laboratory Directed Research & Development Funds @ Sandia
- Eric Vugrin, Casey Glatter, Seth Hanson and Thomas Tarman (SNL) for providing `minimega` and math model data
- Aaron Olson, Mike Eldred and Teresa Portone (SNL) for the invaluable feedback on stochastic noise in sampling estimators

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energys National Nuclear Security Administration under contract DE-NA-0003525.

Why are we interested in network models?

- Network **operators**: understand the potential impacts of changes **before implementing** them
- Network **designers**: understand trade-offs **before** network **creation**

Network modeling refers to:

- **Simulation:** similar to their physics-modeling counterparts and they are **based on a deep understanding of the underlying processes** to simulate network components and interactions in software
- **Emulation:** run the real software on virtualized hardware thus it is able to **capture unknown or not well-understood behaviors**

Examples of Network modeling at Sandia (Courtesy of David Fritz, SAND2018-3927[3])

- **DevOps:** Ensure operation of new hardware, software, services in high-consequence environments. Predictive analysis to detect malfunctions, misconfigurations and malicious consequences
- **Malware:** Understanding of malware through pseudo-in situ execution
- **ICS/SCADA:** Under uncertain threats, what are the best countermeasures for my IT-connected ICS systems? Can we detect attacks? Can we assess resiliency of the IT-controls over the entire power grids?
- **Nuclear Weapons:** Can we assure Communication, Command and Control regardless of network state and threats?

---

[3] http://minimega.org/presentations/gt_2018.slide#7