Security and Limitations of Cyber-Physical Systems

Lecture 3: Defense Mechanisms

ROYAL INSTITUTE
OF TECHNOLOGY

Henrik Sandberg **André Teixeira**

Department of Automatic Control
ACCESS Linnaeus Centre, KTH Royal Institute of Technology
Stockholm, Sweden

Linköping University
August 24-26, 2015

ROYAL INSTITUTE OF TECHNOLOGY

Course Outline

- Monday (8:30-10:30):
  - Lecture 1 (HS): Introduction, data attacks against non-dynamic systems, power network monitoring, security index, graph min cut

- Tuesday
  - 8:30-12:30:
    - Lecture 2 (HS): Attack space for cyber-physical systems: DoS, undetectable, stealth, covert, bias, replay attacks
    - **Lecture 3 (AT): Defense mechanisms, risk management, anomaly detectors, watermarking**
  - 15:30-16:30:
    - Exercise session (Graph min cut, security index)

- Wednesday (8:30-10:30):
  - Lecture 4 (HS): Physical limits of control implementations
  - Exercise session

ROYAL INSTITUTE OF TECHNOLOGY

Key References for Lecture

- [1] André Teixeira, Kin Sou, Henrik Sandberg, and Karl H. Johansson. "Secure Control Systems: A Quantitative Risk Management Approach". IEEE Control System Magazine, vol. 35, no. 1, pp. 24-25, Feb. 2015

- [2] Inseok Hwang, Sungwan Kim, Youdan Kim, C.E. Seah: "A Survey of Fault Detection, Isolation, and Reconfiguration Methods," *Control Systems Technology, IEEE Transactions on* , vol.18, no.3, pp.636,653, May 2010

- [3] Yilin Mo, Sean Weerakkody, Bruno Sinopoli: "Physical Authentication of Control Systems". IEEE Control Systems Magazine, vol. 35, no. 1, pp. 93-109, February 2015.

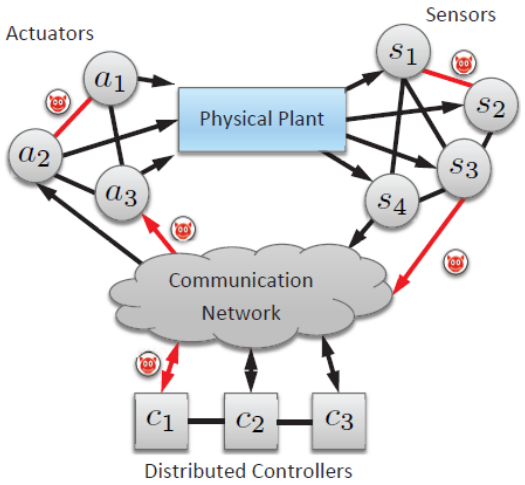
- [4] André Teixeira, Iman Shames, Henrik Sandberg, and Karl H. Johansson, "Revealing Stealthy Attacks in Control Systems". In Proc. 50th Annual Allerton Conference on Communication, Control, and Computing, Allerton, IL, USA, 2012

ROYAL INSTITUTE OF TECHNOLOGY

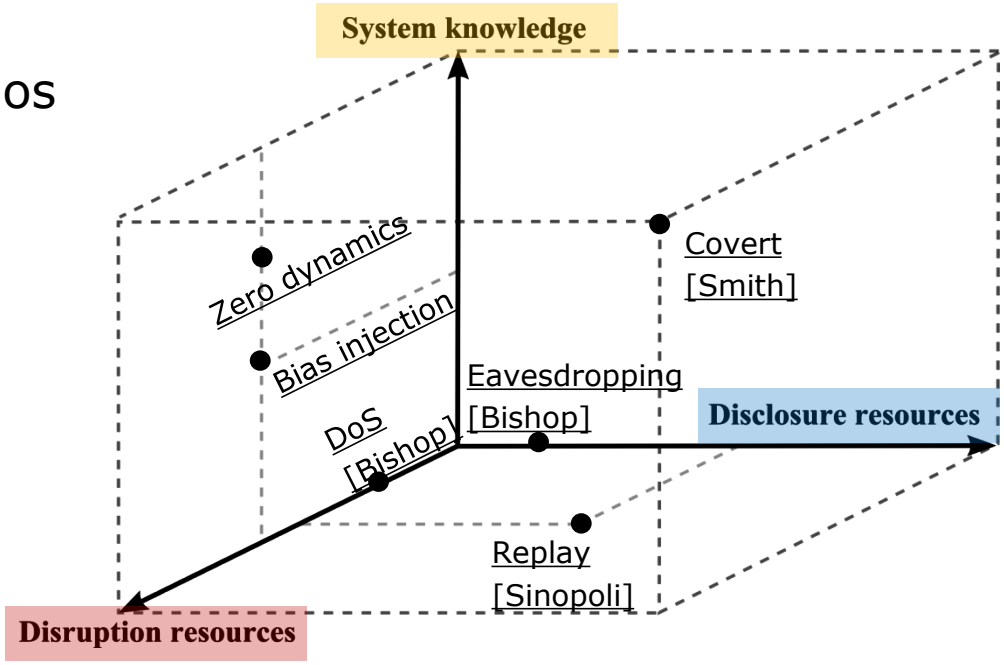
Lecture 3

- Risk management
  - security metrics, quantifying impact [1]
- Anomaly detectors (for detectable attacks) [2]
- Watermarking (against undetectable attacks)
  - Replay attacks [3]
  - Zero dynamics attacks [4]

Motivation



- Complex systems with numerous attack scenarios
- Too costly to secure the entire system against all attack scenarios
- What scenarios to prioritize?
- What components to protect?
- How to mitigate undetectable attacks?



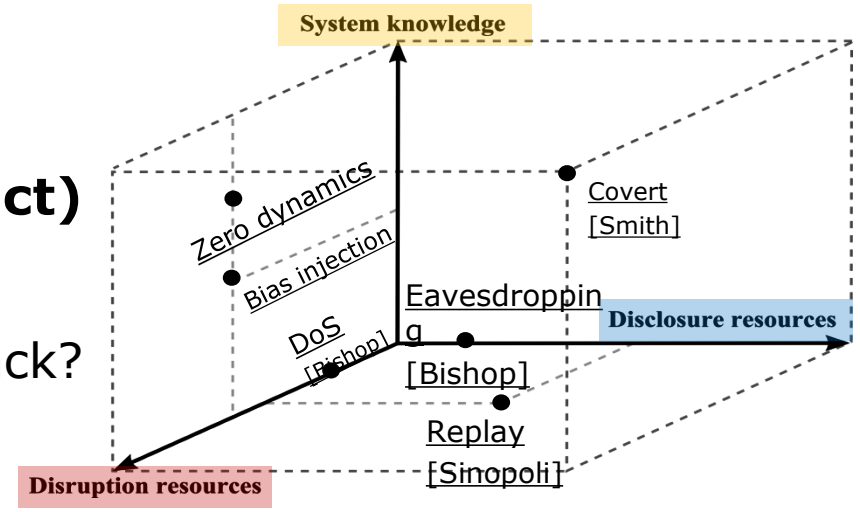
ROYAL INSTITUTE OF TECHNOLOGY

Goals of Lecture 3

- Analyze and compare existing attack scenarios in a risk management framework
- Understand the basics of anomaly detection
- Design methods to reveal undetectable attacks

Defining Risk

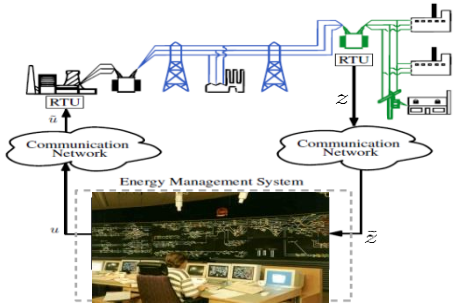
Risk = (Scenario , Likelihood, Impact)



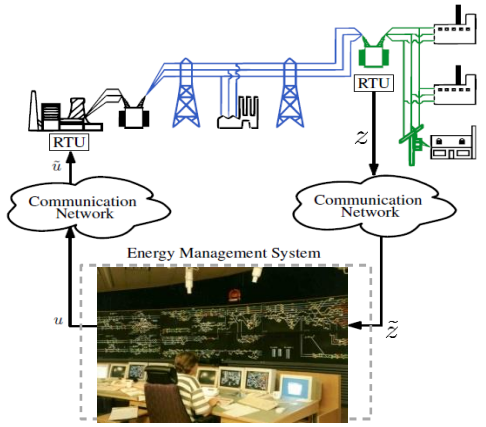
- Scenario
  - How to describe the system under attack?
  - (recall Lecture 2)

- Likelihood
  - How much effort does a given attack require?
  - (compare to security index, Lecture 1)

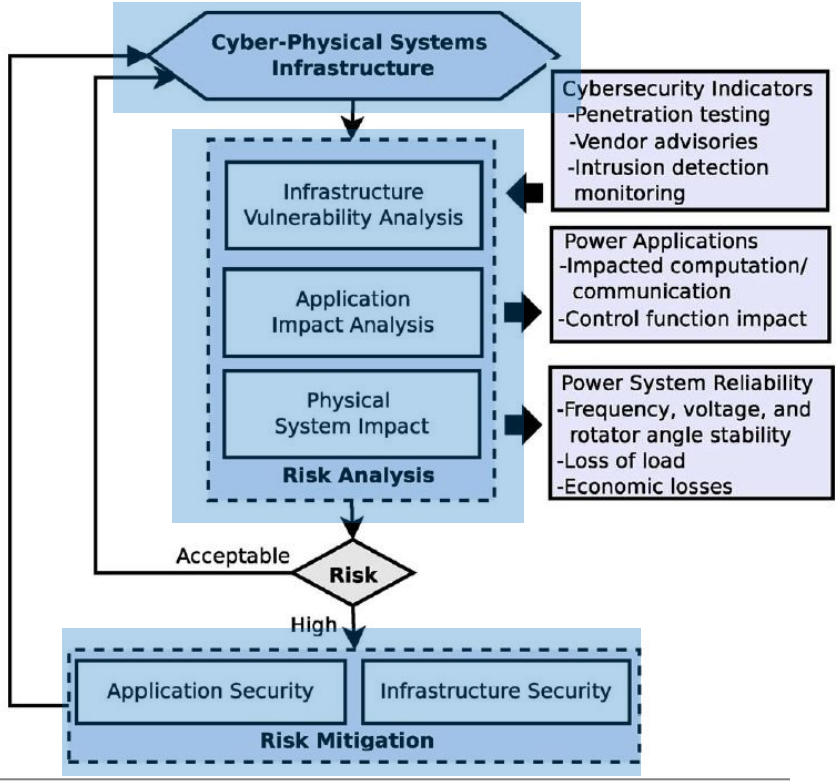
- Impact
  - What are the consequences of an attack?
  - (relate to control objectives)



Risk Management

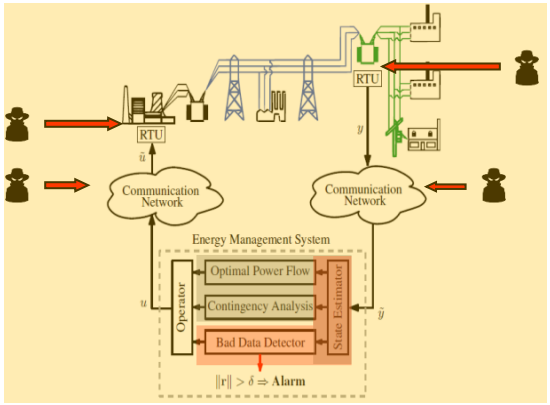


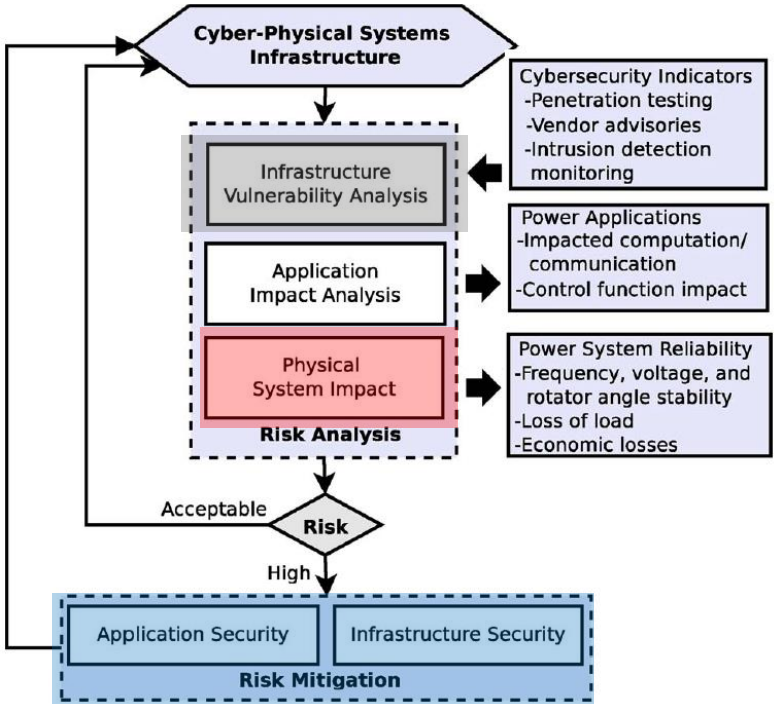
- Main steps in risk management
  - Scope definition
    - Models, Scenarios, objectives
  - Risk Analysis
    - Threat Identification
    - Likelihood Assessment
    - Impact Assessment
  - Risk Treatment
    - Prevention, Detection, Mitigation



Risk Management

- How to model adversaries and attacks?
  - *Lecture 2*
- How to measure likelihood (attack effort)?
  - *Lectures 1 and 2*
- How to compute impact?
  - *Lecture 2*
  - ***This lecture***
- How to design protection and detection mechanisms?
  - ***This lecture***

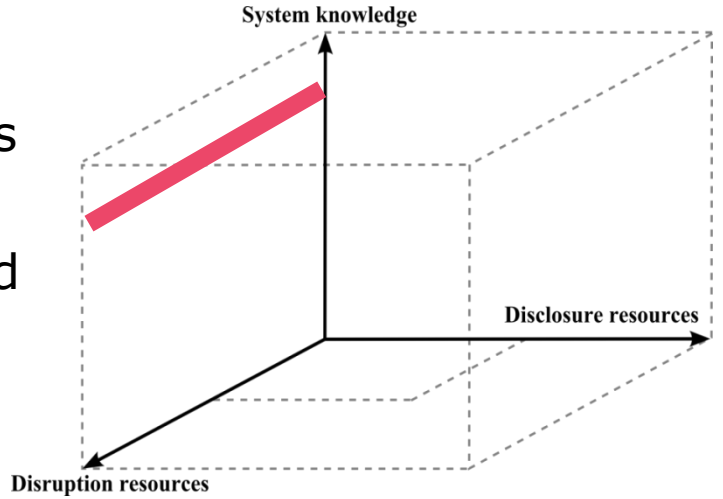




Risk Analysis for Dynamical Systems

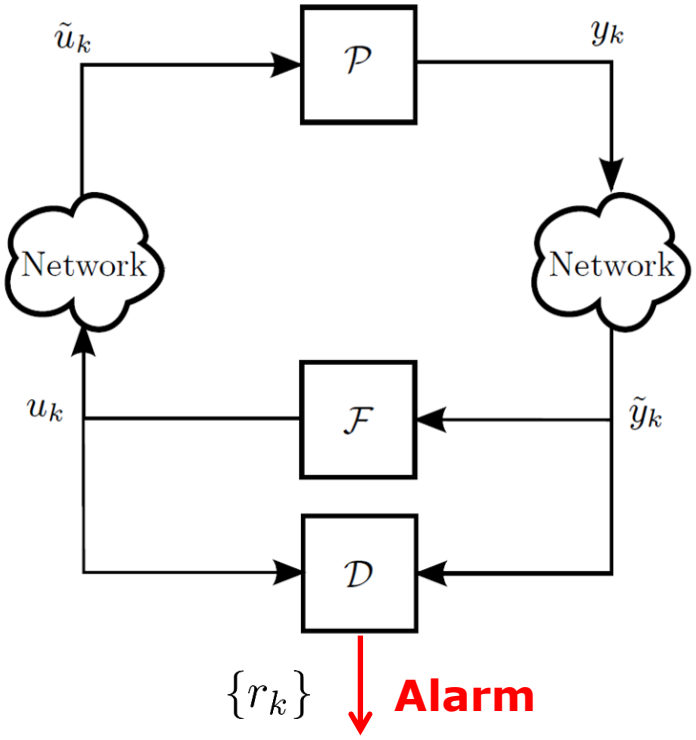
- Analysis of effort and impact of stealthy attacks

- Cases considered here:
  1. Minimum resource attacks
  2. Maximum impact attacks
  3. Maximum impact bounded resource attacks



- Considered attacks are in *open loop*. No disclosure resources explicitly used (works due to linearity of systems)

Networked Control System



- Physical Plant

$$\mathcal{P}: \begin{cases} x_{k+1} = Ax_k + B\tilde{u}_k + Gw_k + Ff_k \\ y_k = Cx_k + v_k \end{cases}$$

- Feedback Controller

$$\mathcal{F}: \begin{cases} z_{k+1} = A_c z_k + B_c \tilde{y}_k \\ u_k = C_c z_k + D_c \tilde{y}_k \end{cases}$$

- Anomaly Detector

$$\mathcal{D}: \begin{cases} s_k = A_e s_k + B_e u_k + K_e \tilde{y}_k \\ r_k = C_e s_k + D_e u_k + E_e \tilde{y}_k \end{cases}$$

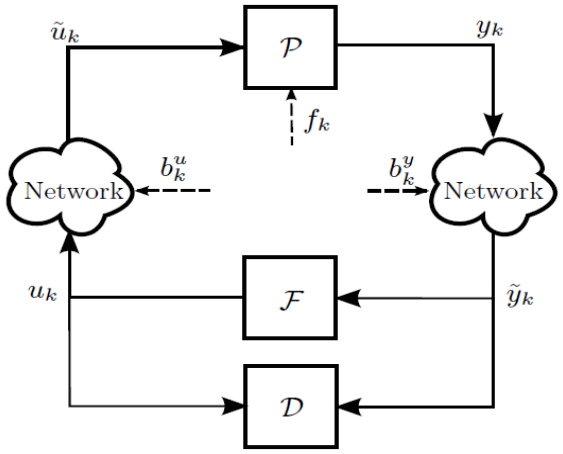
Networked Control System under Attack

- Cyber-Physical Attacks
  - Disclosure Component

$$\mathcal{I}_k := \mathcal{I}_{k-1} \cup \begin{bmatrix} \Upsilon^u & 0 \\ 0 & \Upsilon^y \end{bmatrix} \begin{bmatrix} u_k \\ y_k \end{bmatrix}$$

  - Disruptive component

$$a_k = [f_k^\top \quad b_k^{u\top} \quad b_k^{y\top}]^\top$$



- Closed-loop Dynamics

$$\eta_k = [x_k^\top \quad z_k^\top]^\top$$

$$\eta_{k+1} = \mathbf{A}\eta_k + \mathbf{B}a_k + \mathbf{G}\begin{bmatrix} w_k \\ v_k \end{bmatrix}$$

$$\tilde{y}_k = \mathbf{C}\eta_k + \mathbf{D}a_k + \mathbf{H}\begin{bmatrix} w_k \\ v_k \end{bmatrix}$$

- Anomaly Detector

$$\xi_k = [\eta_k^\top \quad s_k^\top]^\top$$

$$\xi_{k+1} = \mathbf{A}_e\xi_k + \mathbf{B}_e a_k + \mathbf{G}_e\begin{bmatrix} w_k \\ v_k \end{bmatrix}$$

$$r_k = \mathbf{C}_e\xi_k + \mathbf{D}_e a_k + \mathbf{H}_e\begin{bmatrix} w_k \\ v_k \end{bmatrix}$$

  - Alarm triggered if:

$$\|r_k\| \geq \delta_r + \delta_\alpha, \quad \delta_\alpha \in \mathbb{R}^+$$

1. Minimum Resource Attack: Dynamical Case

Dynamical anomaly detector for closed-loop system:

$$\xi_{k+1} = \mathbf{A_e}\xi_k + \mathbf{B_e}a_k + \mathbf{G_e}w_k$$
$$r_k = \mathbf{C_e}\xi_k + \mathbf{D_e}a_k + \mathbf{H_e}v_k$$

Lift to time interval $[0, N]$
with zero-initial conditions and no noise:

$$\underbrace{\begin{bmatrix} r_0 \\ r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}}_{\mathbf{r}} = \underbrace{\begin{bmatrix} \mathbf{D_e} & 0 & \dots & 0 \\ \mathbf{C_eB_e} & \mathbf{D_e} & \dots & 0 \\ \mathbf{C_eA_eB_e} & \mathbf{C_eB_e} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \mathbf{C_eA_e}^{N-1}\mathbf{B_e} & \mathbf{C_eA_e}^{N-2}\mathbf{B_e} & \dots & \mathbf{D_e} \end{bmatrix}}_{\mathcal{T}_r} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}}_{\mathbf{a}}$$

1. Minimum Resource Attack: Dynamical Case

Signal strength channel $\dot{w}\psi$

$$\min_{\mathbf{a}} \|h_p(\mathbf{a})\|_0$$

such that

$$h_p(\mathbf{a}) = [\|\mathbf{a}_{(1)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(i)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(q_a)}\|_{\ell_p}]$$

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

$$\mathbf{a} \in \mathcal{G}$$

- Minimize disruption resources (#channels attacked)
- No alarms (threshold δ_α)
- Reach attack goals $\mathcal{G}$ (compare to security index)

1. Minimum Resource Attack: Formulate as MILP (1)

$\min_{\mathbf{a}} \|h_p(\mathbf{a})\|_0$

$h_p(\mathbf{a}) = [\|\mathbf{a}_{(1)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(i)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(q_a)}\|_{\ell_p}]$
$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$
$\mathbf{a} \in \mathcal{G}$

Note that

$$\|h_p(\mathbf{a})\|_0 \leq \epsilon$$

can equivalently be formulated as

$$\begin{aligned}
\mathbf{a}_{(i)} &\leq M_h \mathbf{z}_i \mathbf{1} \qquad \forall\, i = 1, \ldots,\ q_a \\
-\mathbf{a}_{(i)} &\leq M_h \mathbf{z}_i \mathbf{1} \qquad \forall\, i = 1, \ldots,\ q_a \\
\sum_{i=1}^{q_a} \mathbf{z}_i &\leq \epsilon \\
\mathbf{z}_i &\in \{0, 1\} \qquad \forall\, i = 1, \ldots,\ q_a.
\end{aligned}$$

where M_h is a large constant (“infinity”)

1. Minimum Resource Attack: Formulate as MILP (2)

$$\min_{\mathbf{a},\,\epsilon} \epsilon$$

such that

$$h_p(\mathbf{a}) = [\|\mathbf{a}_{(1)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(i)}\|_{\ell_p}, \ldots, \|\mathbf{a}_{(q_a)}\|_{\ell_p}]$$

$$\|h_p(\mathbf{a})\|_0 \leq \epsilon$$

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

$$\mathbf{a} \in \mathcal{G}$$

- Minimize disruption resources (#channels attacked)
- No alarms (threshold δ_α)
- Reach attack goals $\mathcal{G}$ (compare to security index)

- MILP if $p = q = \infty$

2. Maximum Impact Attack: Dynamical Case

Dynamics of plant and controller:

$$\eta_{k+1} = \mathbf{A}\eta_k + \mathbf{B}a_k + \mathbf{G}w_k$$
$$x_k = \mathbf{C}\eta_k + \mathbf{D}a_k + \mathbf{H}v_k$$

Lifting to time interval $[0, N]$
with zero-initial conditions and no noise:

$$\underbrace{\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} \mathbf{D} & 0 & \dots & 0 \\ \mathbf{CB} & \mathbf{D} & \dots & 0 \\ \mathbf{CAB} & \mathbf{CB} & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ \mathbf{CA}^{N-1}\mathbf{B} & \mathbf{CA}^{N-2}\mathbf{B} & \dots & \mathbf{D} \end{bmatrix}}_{\mathcal{T}_x} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}}_{\mathbf{a}}$$

2. Maximum Impact Attack: Dynamical Case

$$\max_{\mathbf{a}} \|\mathcal{T}_x \mathbf{a}\|_p$$

such that

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

- Maximize impact (push $\|\mathbf{x}\|_p$ far away from equilibrium)
- No alarms (threshold δ_α)

- Not a convex optimization problem!
- Closed-form solution when $p = q = 2$ (use Courant-Fischer Theorem)

See Exercise 6

Sketch on the board!

2. Maximum Impact Attack: Dynamical Case

$$\max_{\mathbf{a}} \|\mathcal{T}_x \mathbf{a}\|_p$$

such that

$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$

Theorem: Bounded solution iff $\ker(\mathcal{T}_r) \subseteq \ker(\mathcal{T}_x)$ **See Exercise 7**

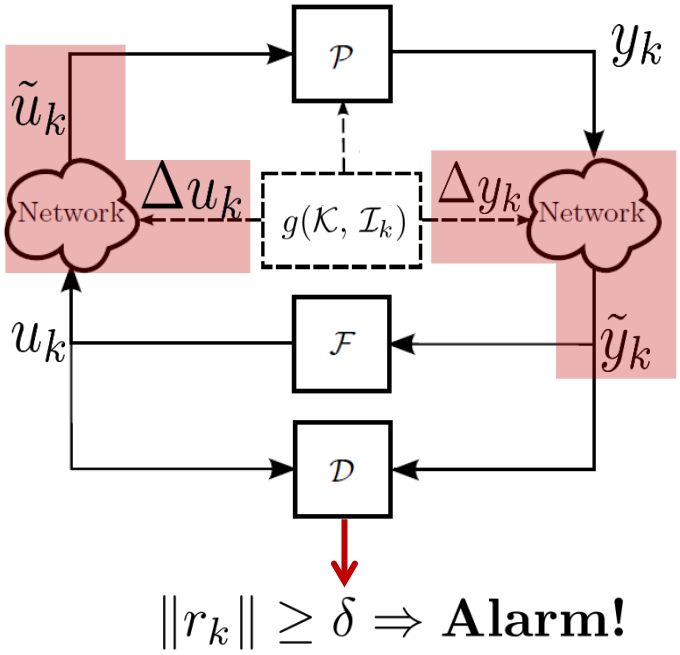
Theorem (*p=q=2*): Assume bounded solution, then

$$\mathbf{a}^\star = \frac{\delta_\alpha}{\|\mathcal{T}_r \mathbf{v}_{\max}\|_2} \mathbf{v}_{\max}, \qquad \|\mathcal{T}_x \mathbf{a}^\star\|_2 = \sqrt{\lambda_{\max}}\delta_\alpha$$

$0 = (\lambda_{\max} \mathcal{T}_r^\top \mathcal{T}_r - \mathcal{T}_x^\top \mathcal{T}_x)\mathbf{v}_{\max}$ ($\lambda_{\max}/\mathbf{v}_{\max}$ max generalized eigenpair)

What happens for infinite time-horizons?

2. Maximum-Impact Attack Infinite Horizon



- Maximum-impact stealthy attack:
  - Maximize "energy" of the state signal
  - Keep the residual signal "small"

$$\begin{aligned} \underset{\{a_k\}_{k=0}^{\infty}}{\text{maximize}} \quad & \sum_{k=0}^{\infty} \|x_k\|_2^2 \\ \text{subject to} \quad & \sum_{k=0}^{\infty} \|r_k\|_2^2 \leq \delta \end{aligned}$$

- If the system has **unstable zero-dynamics**:
  - There exists an *exponentially increasing* input that attains a "small" output

$$\{a_k\}_{k=0}^{\infty} : \; r_k \approx 0, \quad \forall k$$

$$\|a_k\| \to \infty, \quad \|x_k\| \to \infty$$

See Exercise 7

3. Maximum Impact Bounded Resource Attack

$$\max_{\mathbf{a}} \|\mathcal{T}_x \mathbf{a}\|_p$$

such that

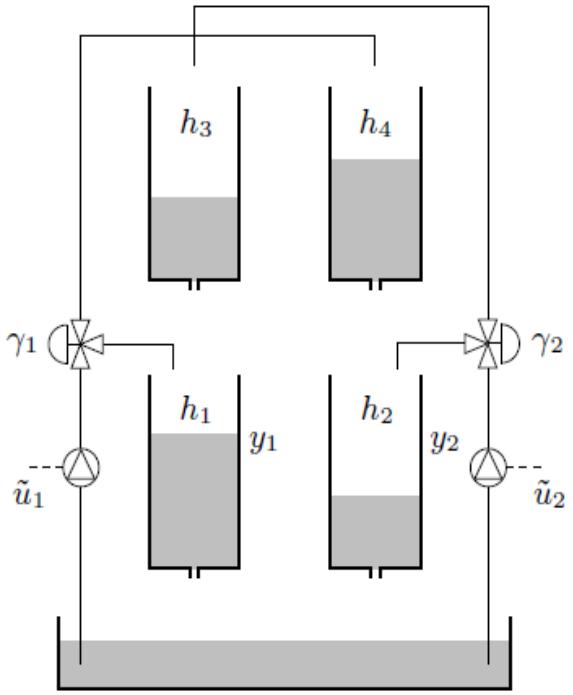
$$\|\mathbf{r}\|_q = \|\mathcal{T}_r \mathbf{a}\|_q \leq \delta_\alpha$$
$$\|h_p(\mathbf{a})\|_0 \leq \epsilon$$

- Maximize impact (push $\|\mathbf{x}\|_p$ far away from equilibrium)
- No alarms (threshold δ_α)
- Use no more than ϵ channels

Jointly considers impact and likelihood

- $p = q = \infty$ can be formulated as MILP (see slide 15)

Numerical Example



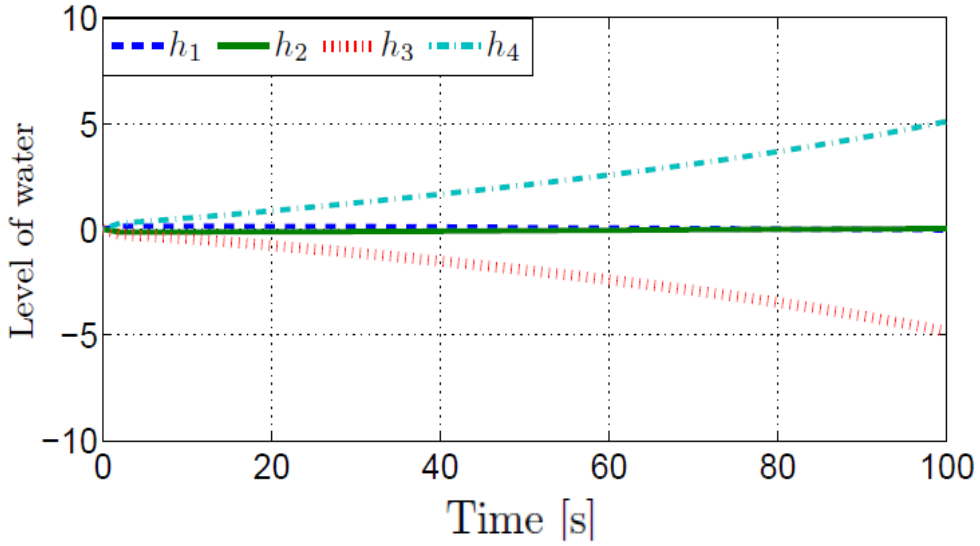
$$\dot{h}_1 = -\frac{a_1}{A_1}\sqrt{2gh_1} + \frac{a_3}{A_1}\sqrt{2gh_3} + \frac{\gamma_1 k_1}{A_1}u_1,$$
$$\dot{h}_2 = -\frac{a_2}{A_2}\sqrt{2gh_2} + \frac{a_4}{A_2}\sqrt{2gh_4} + \frac{\gamma_2 k_2}{A_2}u_2,$$
$$\dot{h}_3 = -\frac{a_3}{A_3}\sqrt{2gh_3} + \frac{(1-\gamma_2)k_2}{A_3}u_2,$$
$$\dot{h}_4 = -\frac{a_4}{A_4}\sqrt{2gh_4} + \frac{(1-\gamma_1)k_1}{A_4}u_1,$$

- Wireless LQG controller
- 4 channels: 2 actuators and 2 measurements
- Minimum phase or non-minimum phase depending on γ_1, γ_2

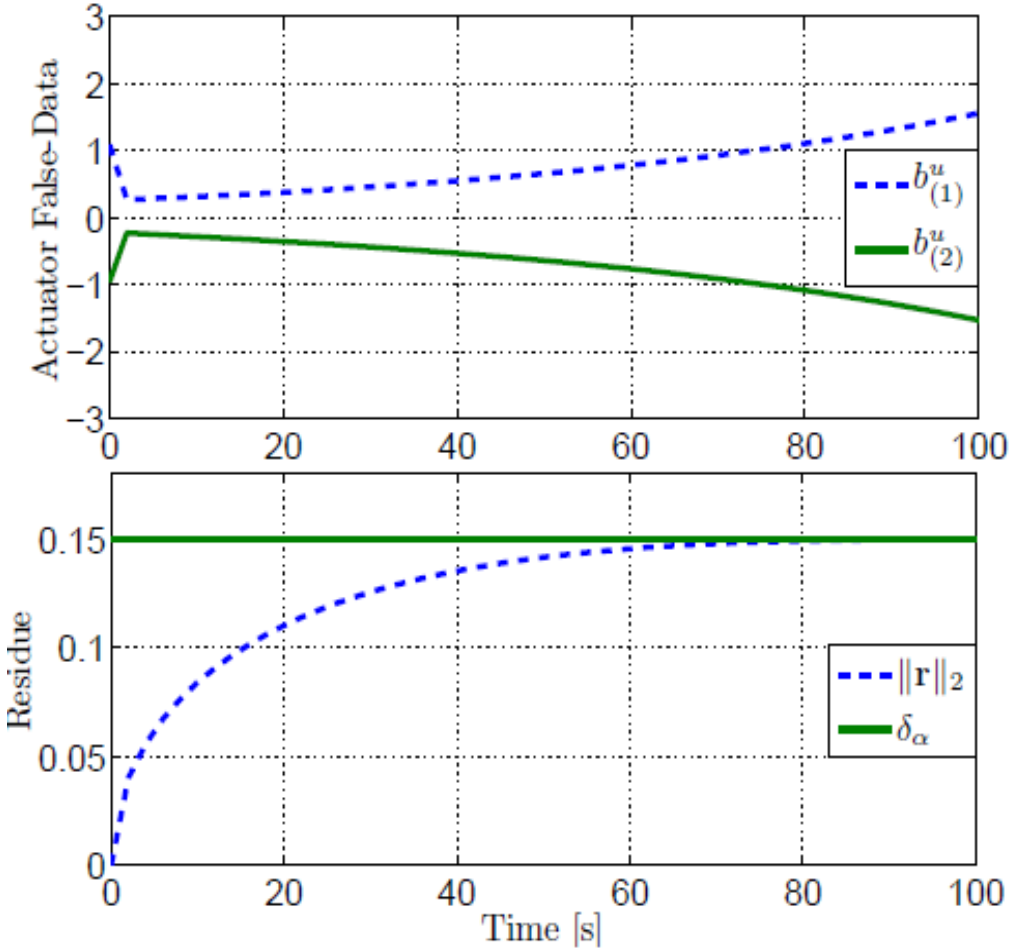
Numerical Example (Non-Min Phase)

Values of $\|\mathbf{x}\|_p$ for maximum impact formulation with $p = q = 2,\ \delta_\alpha = 0.15$

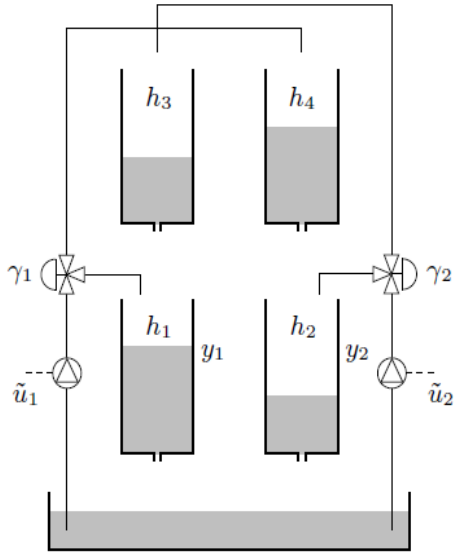
	$\|h_p(\mathbf{a})\|_0$			
	1	**2**	**3**	**4**
Minimum phase	1.15	140.39	∞	∞
Non-minimum phase	2.80	689.43	∞	∞

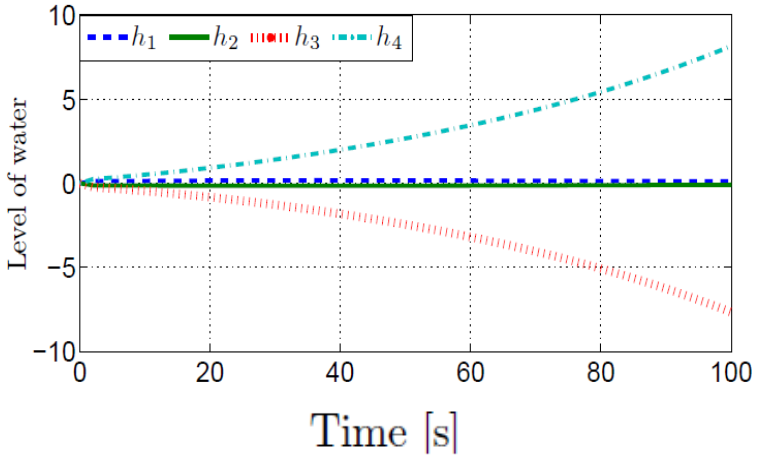


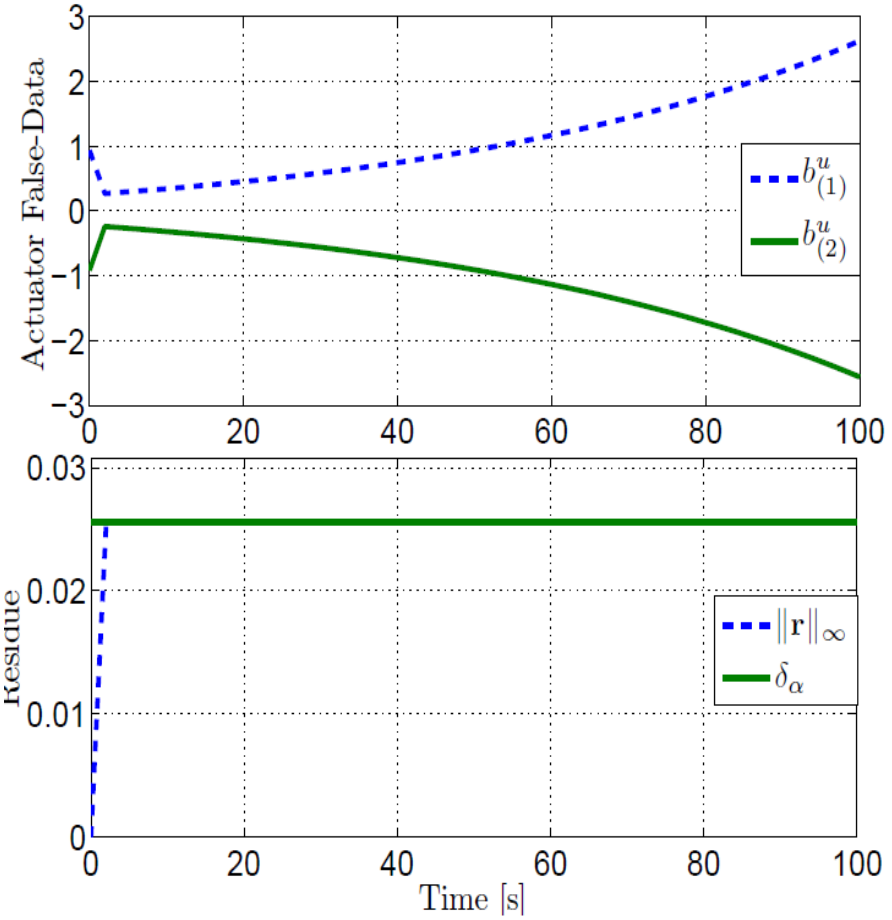
Numerical Example (Non-Min Phase)



Numerical Example (MILP)

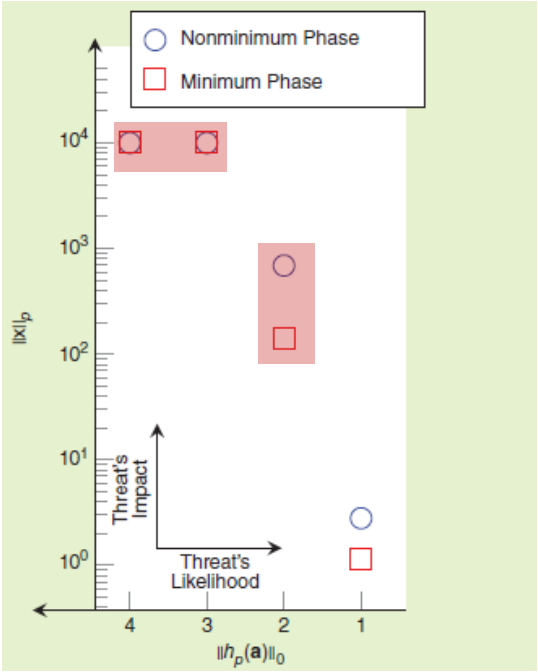






What components to protect?

- Attacks on 3 or more components have very **high impact**
  - Must protect 2 components
- Attacks on 2 compoents still have **high risk**
- What to protect?
  - Protecting ($\{u_1, u_2\}$) yields the **lowest risk**

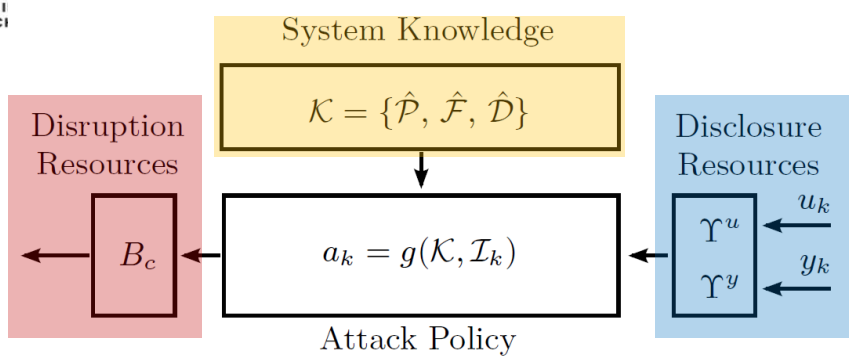


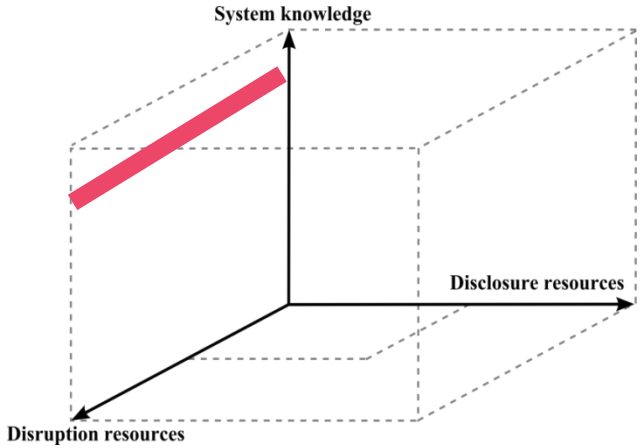
ROYAL INSTITUTE OF TECHNOLOGY

Numerical Example

- Maximum Impact Bounded Resource attack illustrated
- 2 channels allowed: MILP selects the actuators
- 3-4 channels allowed: Unbounded impact (any attack on actuators can be hidden by corrupting 2 measurements)
- Infinity norm criteria ($p = q = \infty$) yields more aggressive attack (bounds saturated)
- Not surprisingly, non-min phase plant more sensitive

Summary





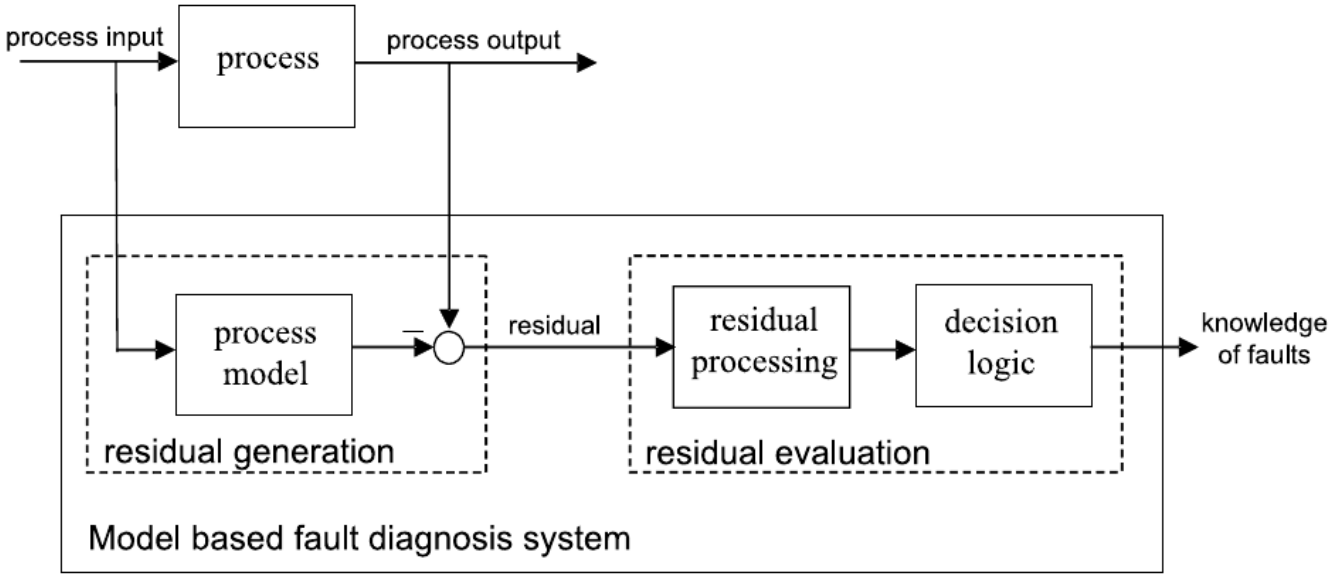
- Tools for quantitative trade-off analysis between attacker's impact and resources: Important for defense prioritization
- For dynamical systems there are *temporal* as well as *spatial (channel) constraints* for attacker to fulfill
  - Enforced through lifting models
- Closed-form solutions and mixed integer linear programming formulations

ROYAL INSTITUTE OF TECHNOLOGY

Lecture 3

- Risk management
  - security metrics, quantifying impact [1]
- Anomaly detectors (for detectable attacks) [2]
- Watermarking (against undetectable attacks)
  - Replay attacks [3]
  - Zero dynamics attacks [4]

Model-Based Fault Diagnosis



- Basic ideas:
  - Compute an expected output (using model information)
  - *Evaluate* the difference between the *real* and *expected* outputs (residual)

Fault-Diagnosis Objectives

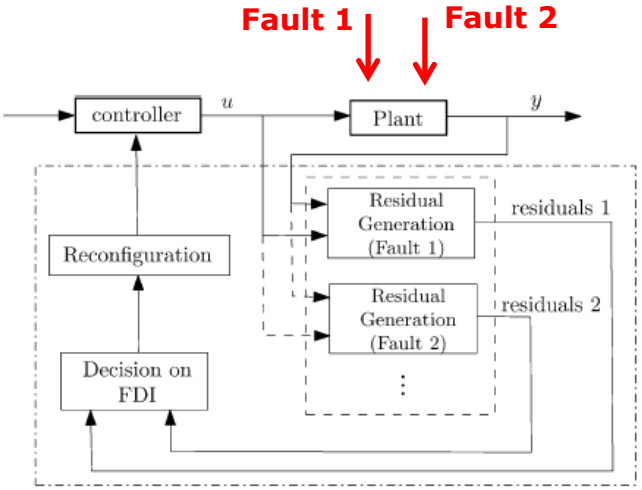
- Fault Detection: detect faults
  - Generate a residual sensitive to faults
  - E.g.: Kalman filter $\hat{x}_{k+1} = A\hat{x}_k + Bu_k + K(y_k - C\hat{x}_k)$

$$r_k = y_k - C\hat{x}_k$$

(no fault)

$$x_k - \hat{x}_k \sim \mathcal{N}(0, P_k)$$
$$r_k \sim \mathcal{N}(0, CP_kC^\top)$$

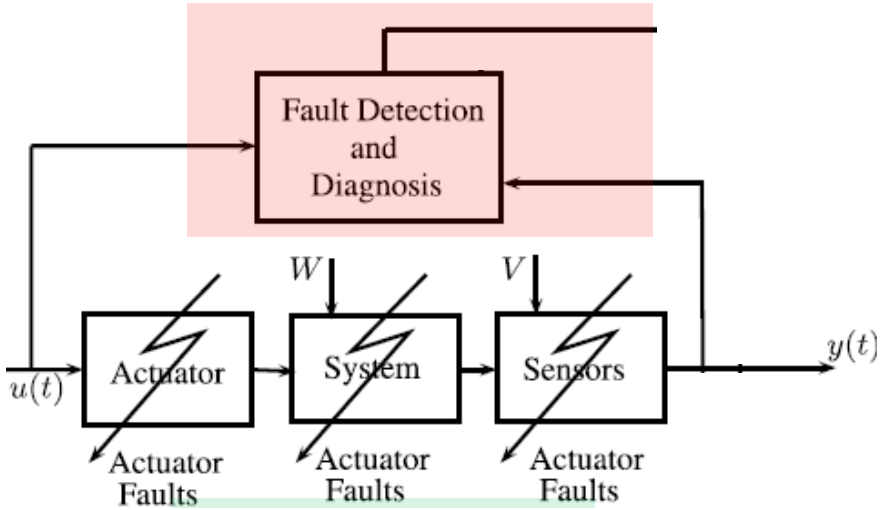
- Fault Isolation: locate the faulty components
  - Generate a set of structured residuals
  - E.g.: $\begin{bmatrix} r_1(s) \\ r_2(s) \end{bmatrix} = \begin{bmatrix} G_{11}(s) & 0 \\ 0 & G_{22}(s) \end{bmatrix} \begin{bmatrix} f_1(s) \\ f_2(s) \end{bmatrix}$



- Fault Identification: estimate the fault signal
  - Use a state estimator
  - Simple example: $\hat{x}_{k+1} - A\hat{x}_k - Bu_k = F\hat{f}_k$

- All **fail** in the presence of **undetectable attacks**

Beyond Fault Diagnosis



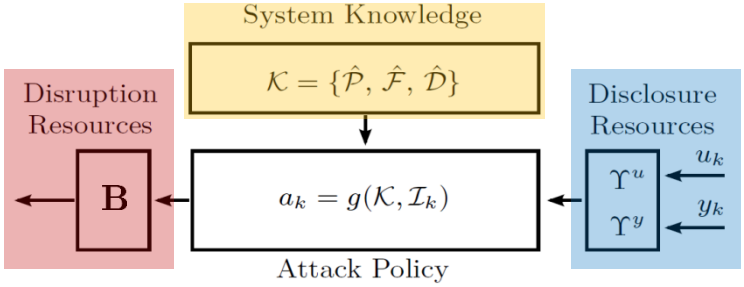
- Often requires redundancy (extra sensors / actuators)
- See [2] for more details and references

ROYAL INSTITUTE OF TECHNOLOGY

Lecture 3

- Risk management
  - security metrics, quantifying impact [1]
- Anomaly detectors (for detectable attacks) [2]
- Watermarking (against undetectable attacks)
  - Replay attacks [3]
  - Zero dynamics attacks [4]

Replay Attack – Phase II [3]



- No more data is recorded

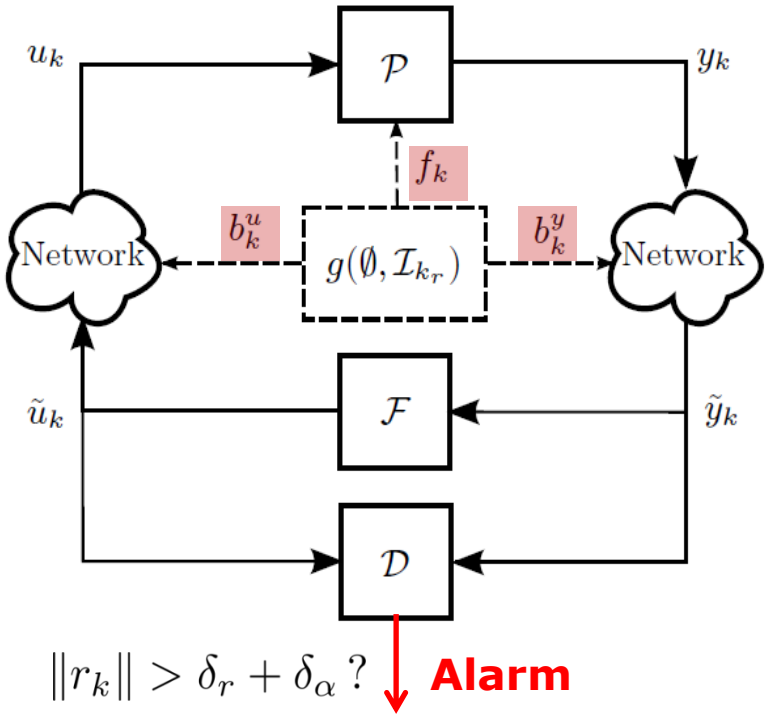
$$\mathcal{I}_k := \mathcal{I}_{k-1}, \quad k > r$$

- The previously recorded data is replayed

$$\begin{bmatrix} b_k^u \\ b_k^y \end{bmatrix} = \begin{bmatrix} \Upsilon^u(u_{k-T} - u_k) \\ \Upsilon^y(y_{k-T} - y_k) \end{bmatrix}$$

- Physical attack is also performed

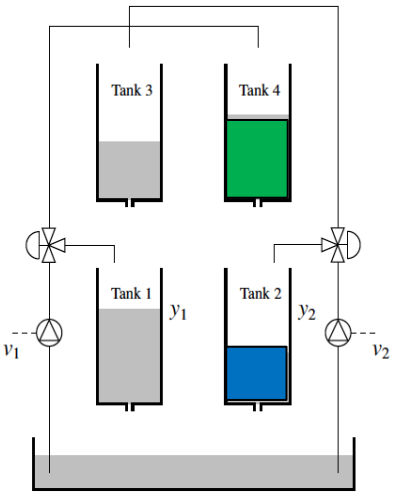
$$f_k = g_f(\emptyset, \mathcal{I}_{k_r})$$

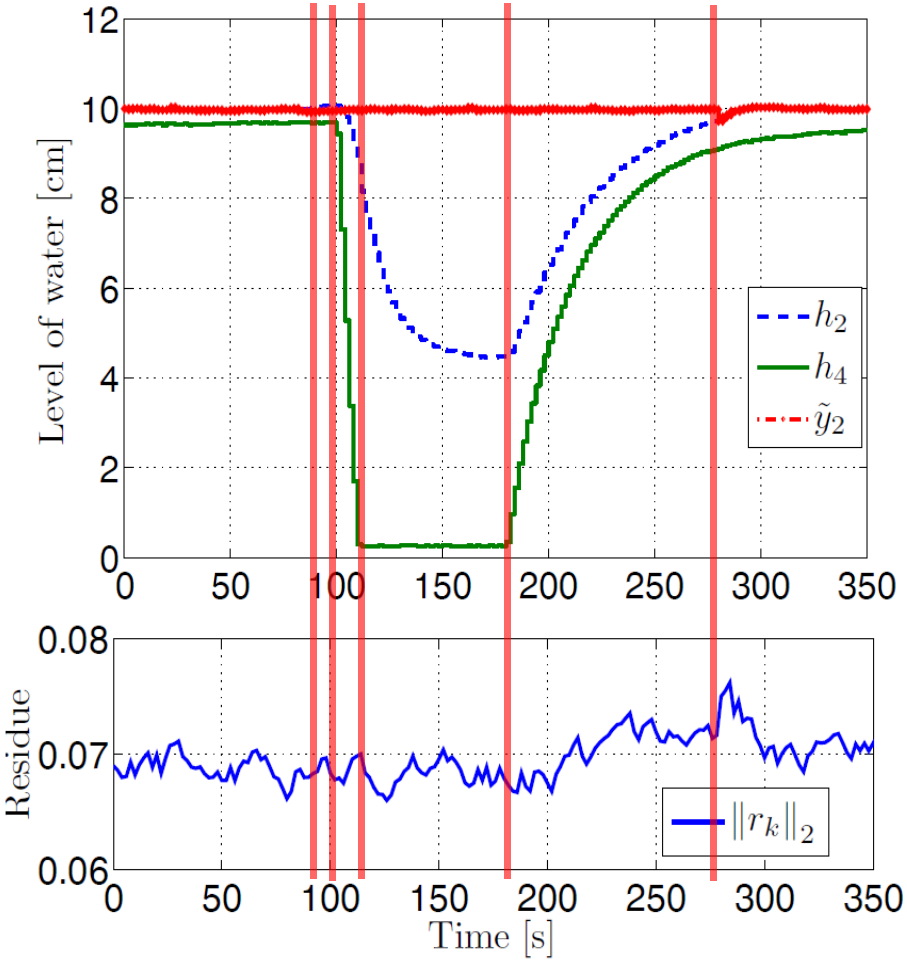


- *No system knowledge is needed:* $\mathcal{K} = \{\hat{P}, \hat{F}, \hat{D}\} = \emptyset$

[Mo and Sinopoli, Allerton, 2009]

Replay Attack - Experiment





- **Attack Goal:** Empty tank 4
- Replay attack on sensor 2
- Physical attack on tank 4
- Tank 4 is **emptied**
- Physical attack ends at t=180s
- Replay attack ends at t=280s
- The attack is **not detected**
  - **Why is it undetectable?**
  - **Can we "make" it detectable?**

Sensor Replay Attack – Analysis [3]

- Residual generated by a Kalman Filter (+LQG controller $u_k = L\hat{x}_k$)

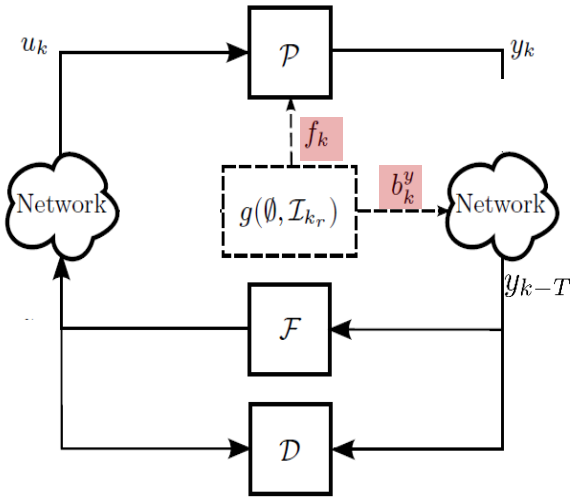
$$\begin{aligned}\hat{x}_{k+1|k} &= A\hat{x}_k + Bu_k \\ \hat{x}_k &= \hat{x}_{k|k-1} + K(y_k - C\hat{x}_{k|k-1}) \\ r_k &= y_k - C\hat{x}_k\end{aligned}$$

(no fault)

$$\begin{aligned}x_k - \hat{x}_k &\sim \mathcal{N}(0, P_k) \\ r_k &\sim \mathcal{N}(0, CP_kC^\top + R)\end{aligned}$$

- Model data replay as a virtual time-shifted plant

$$x_k^v = x_{k-T}, \quad y_k^v = y_{k-T}, \quad r_k^v = r_{k-T}$$



- Residual under sensor replay attack

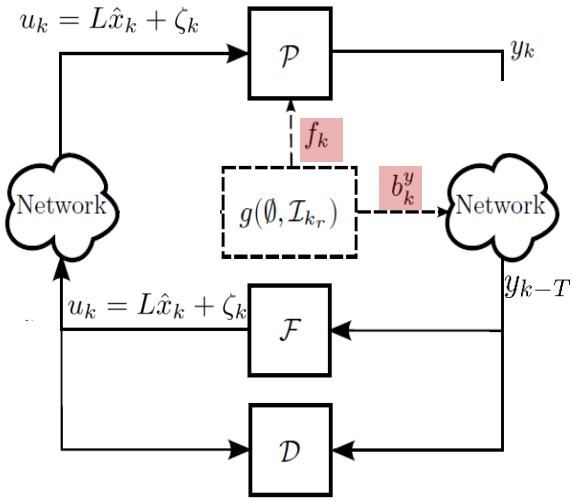
$$r_{k+1} = r_{k+1}^v - C\mathcal{A}(\hat{x}_{0|-1} - \hat{x}_{0|-1}^v)$$

$$\mathcal{A} = (A + BL)(I - KC)$$

- Attack is stealthy if $\mathcal{A}$ is stable
  - **Attacked residual** converges to **healthy residual** $r_k \to r_k^v \sim \mathcal{N}(0, CP_kC^\top + R)$
  - **Relies** on the (virtual) plant and Kalman filter having **the same control policy**

Sensor Replay Attack – Watermarking [3]

- The plant **proactively changes** the control policy
  - Adds noise to control input: $u_k = L\hat{x}_k + \zeta_k$
  - Noise is randomly generated, but **known**



- Residual under sensor replay attack

$$r_{k+1} = r^v_{k+1} - C\mathcal{A}(\hat{x}_{0|-1} - \hat{x}^v_{0|-1}) - C\sum_{i=0}^{k} \mathcal{A}^{k-i} B(\zeta_i - \zeta^v_i)$$

$$\mathcal{A} = (A + BL)(I - KC)$$

- Distribution of residual under attack changes
  - Nominal distribution: $r_k \sim \mathcal{N}(0, CP_kC^\top + R)$
  - Distribution with attack: $r_k \sim \mathcal{N}(\mu_{k-1}, CP_kC^\top + R + \Sigma)$

- Attack can be detected by comparing the two distributions
  - See **Kullback-Leibler divergence** and **Neyman-Pearson test**
  - Detection enabled by **asymmetries** between the time-shifted and real-time models

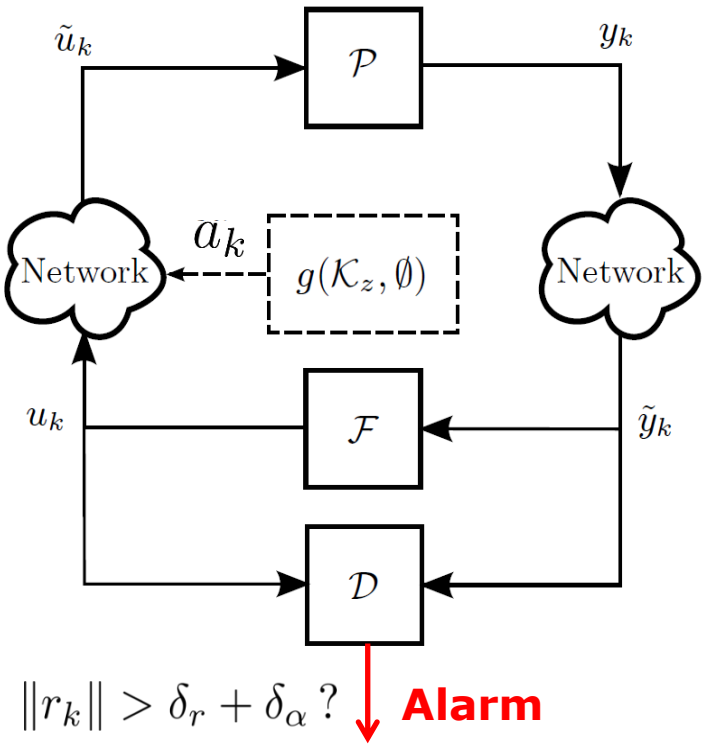
Can we use other forms of asymmetry to detect attacks?

ROYAL INSTITUTE OF TECHNOLOGY

Lecture 3

- Risk management
  - security metrics, quantifying impact [1]
- Anomaly detectors (for detectable attacks) [2]
- Watermarking (against undetectable attacks)
  - Replay attacks [3]
  - Zero dynamics attacks [4]

Zero-Dynamics Attack Model

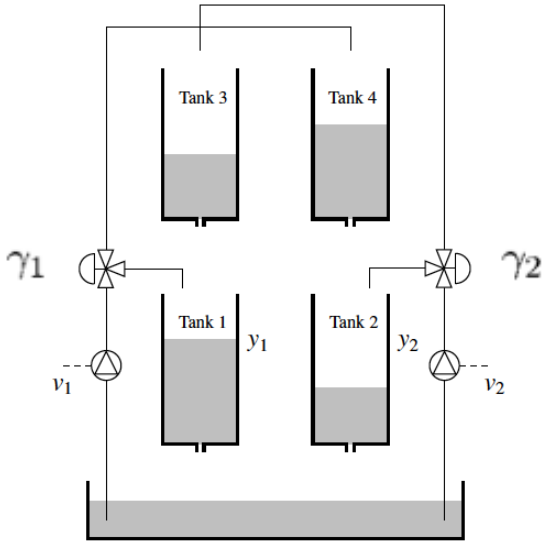


- Physical Plant under attack

$$\mathcal{P}^a : \begin{cases} x_{k+1} = Ax_k + Ba_k \\ y_k = Cx_k \end{cases}, \quad x_0 = 0$$

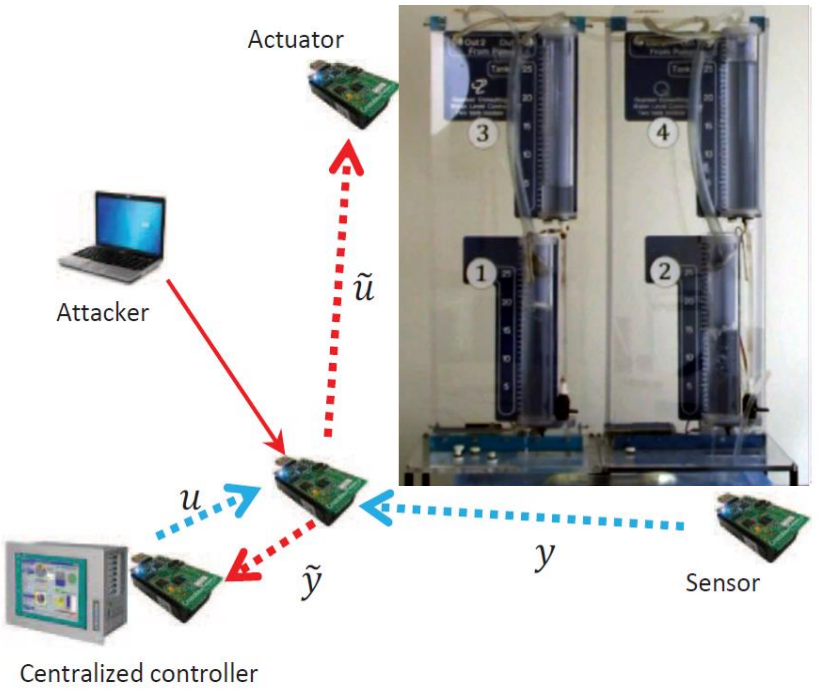
- Attack policy
  - Computed using A, B, C
  - Open-loop policy
- Attack Goals and Constraints
  - Reach an unsafe state
  - Remain stealthy

Testbed for Networked Control System Security



$$\frac{dx}{dt} = \begin{bmatrix} -\frac{1}{T_1} & 0 & \frac{A_3}{A_1 T_3} & 0 \\ 0 & -\frac{1}{T_2} & 0 & \frac{A_4}{A_2 T_4} \\ 0 & 0 & -\frac{1}{T_3} & 0 \\ 0 & 0 & 0 & -\frac{1}{T_4} \end{bmatrix} x + \begin{bmatrix} \frac{\gamma_1 k_1}{A_1} & 0 \\ 0 & \frac{\gamma_2 k_2}{A_2} \\ 0 & \frac{(1-\gamma_2) k_2}{A_3} \\ \frac{(1-\gamma_1) k_1}{A_4} & 0 \end{bmatrix} u$$

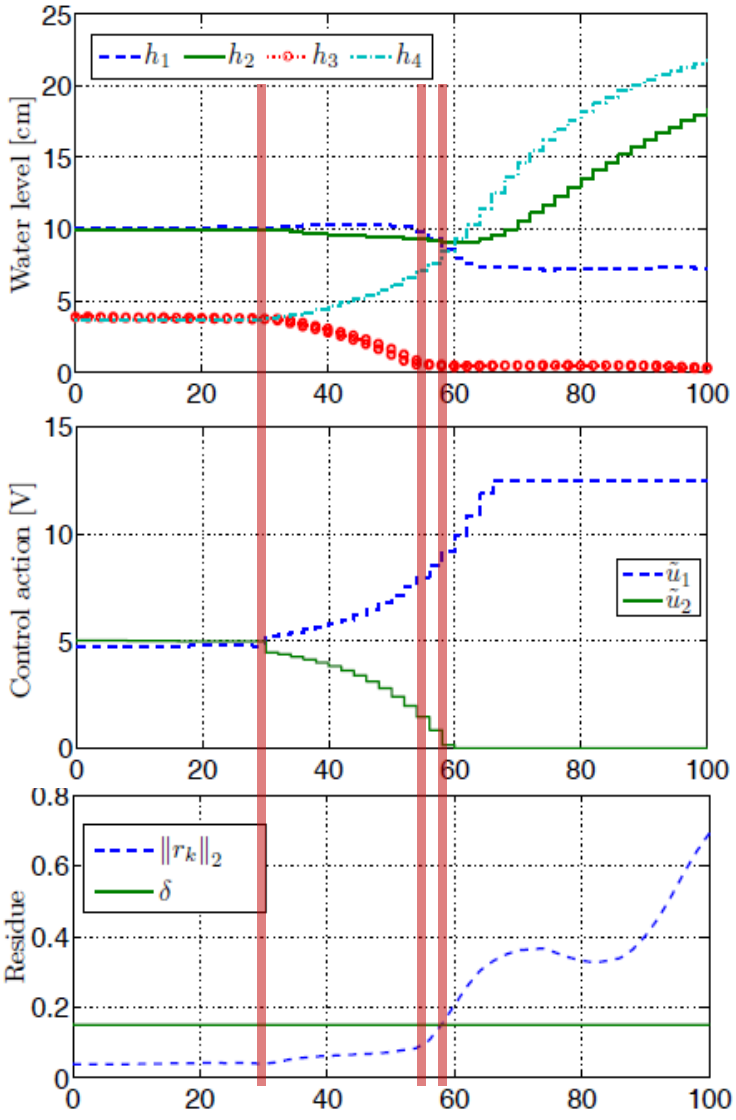
$$y = \begin{bmatrix} k_c & 0 & 0 & 0 \\ 0 & k_c & 0 & 0 \end{bmatrix} x$$

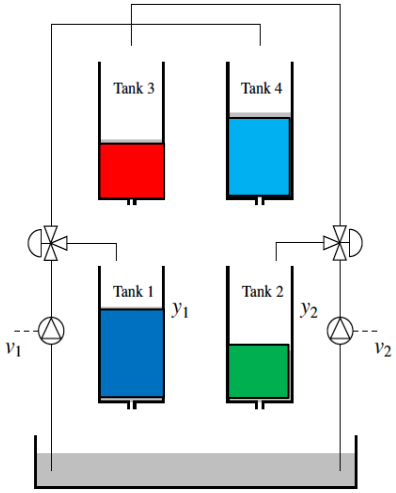


Quadruple-tank process has an unstable zero if $0 < \gamma_1 + \gamma_2 < 1$

[Johansson, IEEE TCST, 2000]

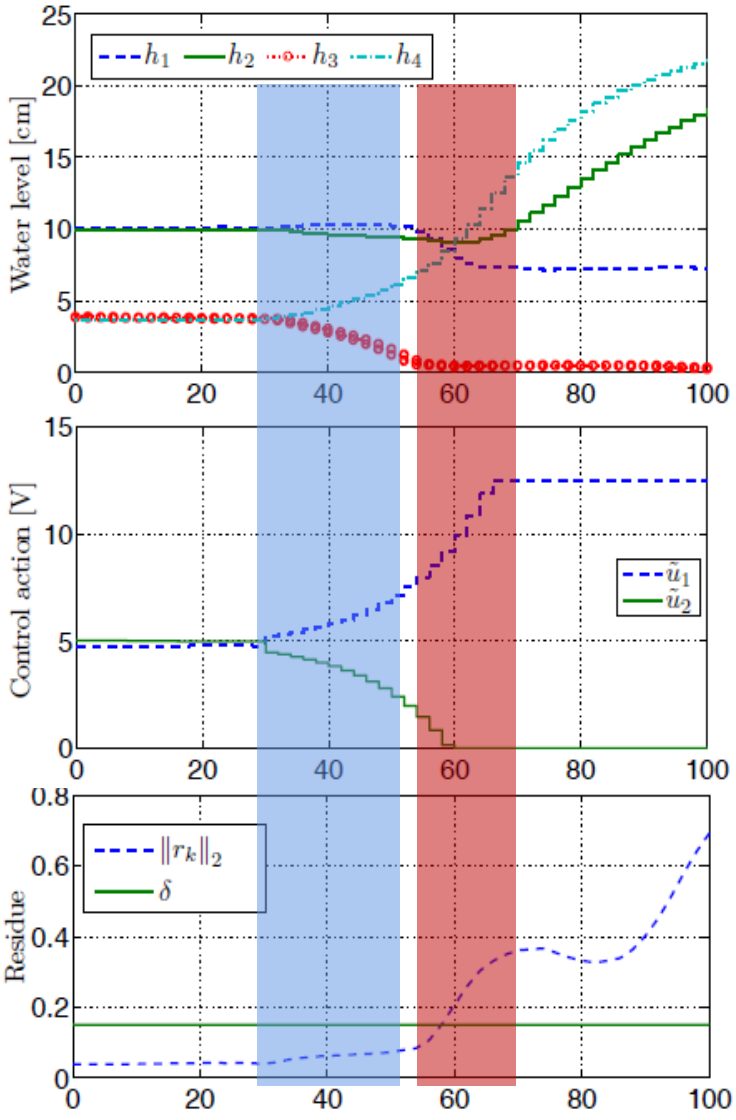
Experimental Result (Lecture 2)

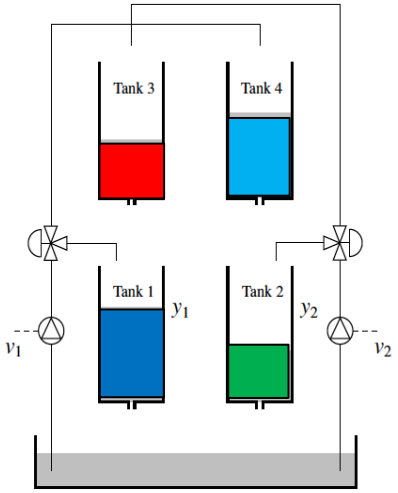




- **Attack Goal:** Empty tank 3
- Zero-dynamics attack on both actuators
- Tank 3 becomes empty
- The attack is **detected**

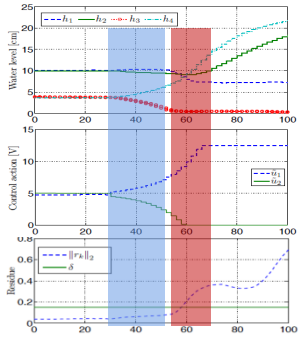
Experimental Result – Why?





- Smooth increase
  - What causes it?
  - Does it compromise the attack's stealthiness?
- Abrupt increase
  - How can it be induced so that attacks are revealed?

Revealing Zero Dynamics Attacks [4]



- Revisit zero dynamics
- Output behavior with initial condition mismatch
- Revealing zero dynamics attacks

Zero Dynamics

- Physical plant

$$\mathcal{P}: \begin{cases} x_{k+1} = Ax_k + Ba_k \\ y_k = Cx_k \end{cases}$$

- Output-zeroing problem:
  - Find z_0 and F such that

$$\begin{matrix} x_{k+1} = (A+BF)x_k \\ 0 = Cx_k \end{matrix}, \quad x_0 = z_0$$

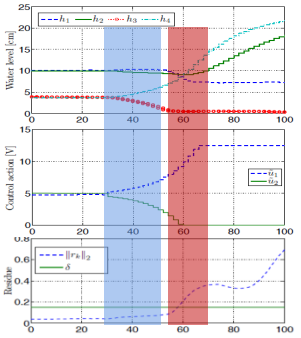
Zero Dynamics

$$x_{k+1} = (A+BF)x_k$$
$$0 = Cx_k,$$

with $x_0 \in \mathcal{V}^\star \subset \ker(\mathrm{C})$ and F such that $(A+BF)\mathcal{V}^\star \subseteq \mathcal{V}^\star$.

Interpretation for Linear Systems?

Zero Dynamics Attack



- Physical plant under attack

$$\mathcal{P}: \begin{cases} x_{k+1} = Ax_k + Ba_k \\ y_k = Cx_k \end{cases}, \quad x_0 = 0$$

- Undetectable attacks

$$\begin{aligned} x_{k+1} &= Ax_k + Ba_k \\ 0 &= Cx_k \end{aligned}, \quad x_0 = 0$$

- Zero dynamics attack policy

$$\begin{aligned} z_{k+1} &= (A + BF)z_k \\ a_k &= Fz_k, \end{aligned}$$

with $z_0 \in \mathcal{V}^\star \subset \ker(\mathrm{C})$ and
F such that $(A + BF)\mathcal{V}^\star \subseteq \mathcal{V}^\star$.

$$\begin{aligned} x_{k+1} &= Ax_k + Ba_k \\ 0 &= Cx_k \end{aligned}, \quad \text{if } x_0 = z_0 \in \mathcal{V}^\star$$

What happens when $x_0 \neq z_0$?

Initial Condition Mismatch

Theorem 1. The output produced by the zero dynamics attack

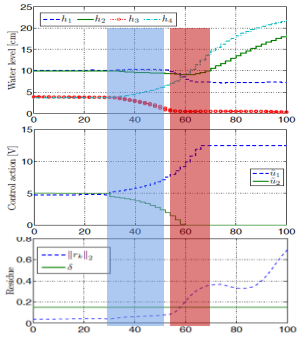
$$\begin{aligned} z_{k+1} &= (A+BF)z_k \\ a_k &= Fz_k \end{aligned}, \quad z_0 \in \mathcal{V}^\star$$

is described by

$$\begin{aligned} e_{k+1} &= Ae_k \\ y_k &= Ce_k \end{aligned}, \quad e_0 = -z_0$$

- Attack is not undetectable if z_0 belongs to the observable subspace of (A, C) (i.e. z_0 yields a zero output)
- If A is stable:
  - the resulting output energy is finite;
  - the output can be made arbitrarily small by scaling down the initial condition z_0

Revealing Zero Dynamics Attacks



Definition: A zero dynamics attack is revealed if $y_k \neq 0$

- Proposed approach (watermarking):
  - change the system dynamics from $\Sigma = (A, B, C)$ to $\tilde{\Sigma} = (\tilde{A}, \tilde{B}, \tilde{C})$

Zero-dynamics attacks are stealthy with respect to the system

$$\begin{bmatrix} x_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} A & BF \\ 0 & A+BF \end{bmatrix} \begin{bmatrix} x_k \\ z_k \end{bmatrix}$$

$$y_k = \begin{bmatrix} C & 0 \end{bmatrix} \begin{bmatrix} x_k \\ z_k \end{bmatrix}$$

for all $x_0 = z_0 \in \mathcal{V}^\star$

Every zero-dynamics attack is revealed if the system

$$\begin{bmatrix} x_{k+1} \\ z_{k+1} \end{bmatrix} = \begin{bmatrix} \tilde{A} & \tilde{B}F \\ 0 & A+BF \end{bmatrix} \begin{bmatrix} x_k \\ z_k \end{bmatrix}$$

$$y_k = \begin{bmatrix} \tilde{C} & 0 \end{bmatrix} \begin{bmatrix} x_k \\ z_k \end{bmatrix}$$

is observable for all $x_0 = z_0 \in \mathcal{V}^\star$

Modifying the input matrix B (1)

- Consider $\tilde{B} = BW$
- **Observation:** attacks remain ***undetectable*** w.r.t to W if and only if the unobservable trajectories are ***not perturbed***

Proof sketch: Check the conditions when $x_0 = z_0 \in \mathcal{V}^\star$ is unobservable:

$$\begin{bmatrix} \lambda I - A & -BF - B(W-I)F \\ 0 & \lambda I - (A+BF) \\ C & 0 \end{bmatrix} \begin{bmatrix} z_0 \\ z_0 \end{bmatrix} = 0 \quad \text{, where} \quad (\lambda I - (A+BF))\, z_0 = 0$$

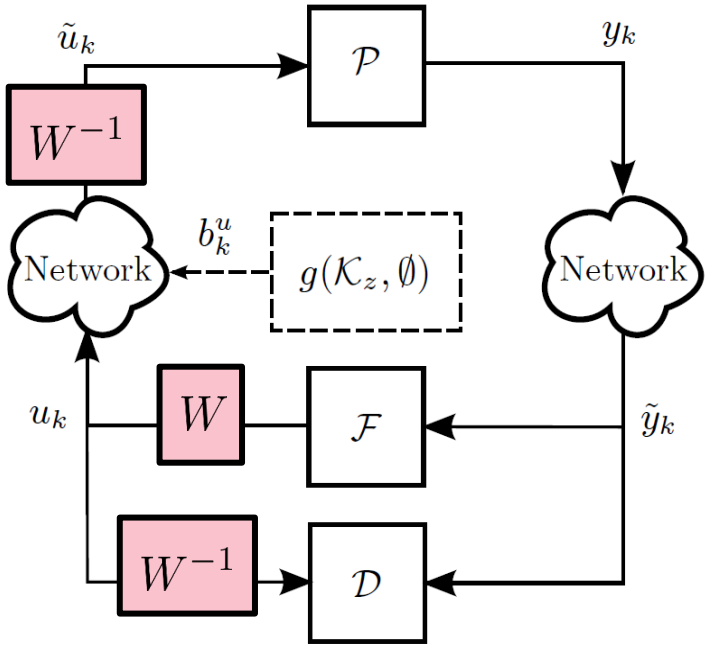
$$B(W-I)Fz_0 = 0$$

- **Revealing attacks:** Choose W such that $B(W-I)Fz_0 \neq 0$

Theorem. All the zero dynamics attacks associated with a given $z_0 \in \mathcal{V}^\star$ remain stealthy with respect to $\tilde{\Sigma} = (A, \tilde{B}, C)$ if and only if $\mathcal{V}^\star \subseteq \ker(B(W-I)F)$

Modifying the input matrix B (2)

- Simply changing B to $\tilde{B}$ affects the system performance under **no attack**
- Coordinated input scaling:
  - Similar to encryption

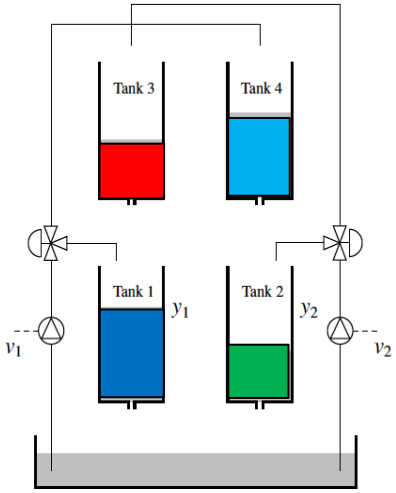


Theorem. Let $z_k = x_k = z \in \mathcal{V}^\star$ and apply $W = \alpha I$ at time k. The output trajectory is described by

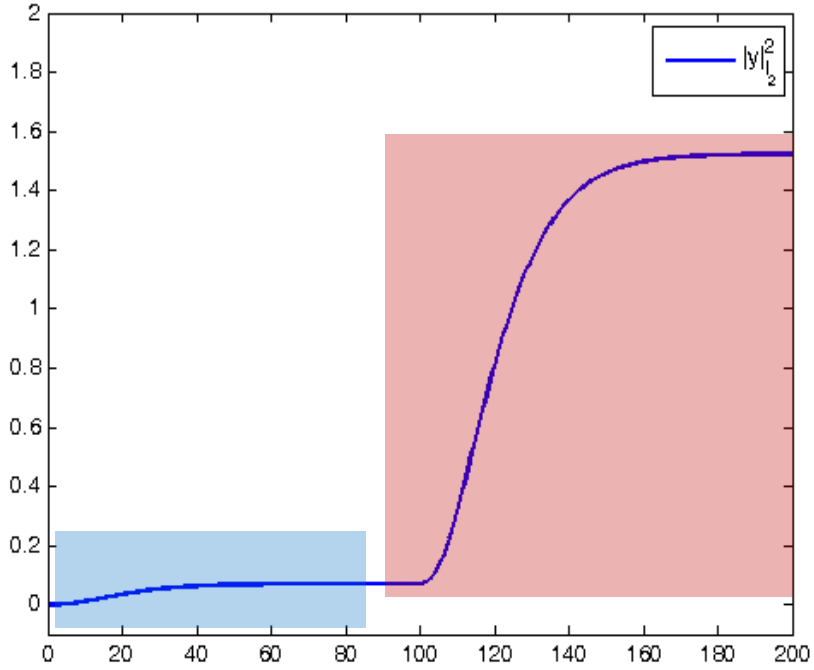
$$\begin{aligned} e_{k+1} &= Ae_k \\ y_k &= Ce_k \end{aligned}, \quad e_0 = (1-\alpha)z$$

- **Revealing attack:** choose α such that y_k is "large" enough.
  - Does not affect the system dynamics

Example – modifying B



- Solution to reveal attacks: input scaling $W = \alpha I$
- Example: choose $\alpha = 0.987$



- Attack begins at $k = 0$
  - Initial condition mismatch
- Input scaling applied at $k = 100$
  - the attack is revealed
- Stable A results in finite output energy

Modifying the output matrix C

- Consider $\tilde{C} = \begin{bmatrix} C \\ \Delta C \end{bmatrix}$
- **Observation:** attacks remain ***undetectable*** w.r.t to $\tilde{C}$ if and only if $\Sigma = (A, B, C)$ and $\tilde{\Sigma} = (A, B, \tilde{C})$ share common unobservable trajectories
- **Revealing attacks:** add measurements so that $\mathcal{X} = \tilde{\mathcal{V}}^\star \cap \mathcal{V}^\star$ becomes empty
  - system dynamics are not affected
  - Requires at most $\dim(\mathcal{V}^\star)$ new measurements

Theorem. There exists a $z_0 \in \mathcal{V}^\star$ generating an undetectable attack to $\tilde{\Sigma} = (A, B, \tilde{C})$ if and only if $(A + BF)$ has an eigenvector in $\mathcal{V}^\star \cap \ker \tilde{C}$

Modifying the system matrix A

- Consider $\tilde{A} = A + \Delta A$
- **Observation:** attacks remain ***undetectable*** w.r.t to $\tilde{A}$ if and only if the unobservable trajectories are ***not perturbed*** (similar to $\tilde{B}$)

Proof sketch: Check the conditions for which $x_0 = z_0 \in \mathcal{V}^\star$ is unobservable:

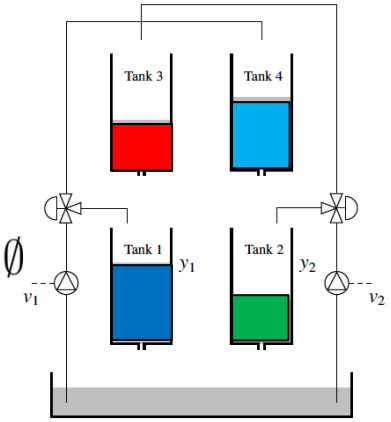
$$\begin{bmatrix} \lambda I - A - \Delta A & -BF \\ 0 & \lambda I - (A + BF) \\ C & 0 \end{bmatrix} \begin{bmatrix} z_0 \\ z_0 \end{bmatrix} = 0 \quad \text{, where } (\lambda I - (A + BF))\, z_0 = 0$$

- $\Delta A z_0 = 0$

- **Revealing attacks:** choose ΔA such that $\mathcal{V}^\star \cap \ker(\Delta A) = \emptyset$
  - Affects the system dynamics and may also require re-designing the controller

Theorem. All the zero-dynamics attacks associated with a given $z_0 \in \mathcal{V}^\star$ remain stealthy with respect to $\tilde{\Sigma} = (\tilde{A}, B, C)$ if and only if $\mathcal{V}^\star \subseteq \ker(\Delta A)$

Example – modifying A

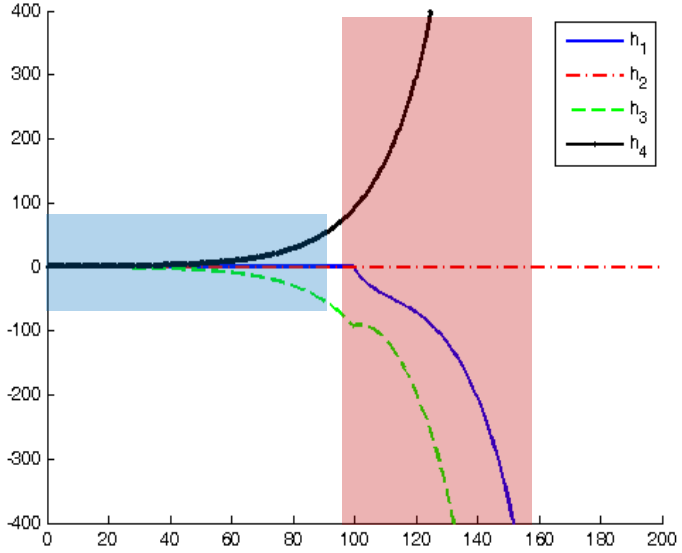


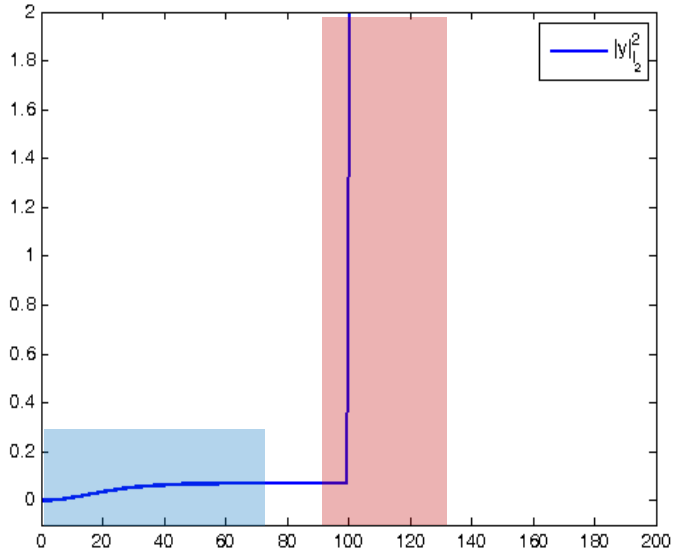
- Solution to reveal attacks: ΔA such that $\mathcal{V}^\star \cap \ker(\Delta A) = \emptyset$

$$V^\star = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Rightarrow \quad \Delta A = \begin{bmatrix} 0 & \Delta \end{bmatrix}$$

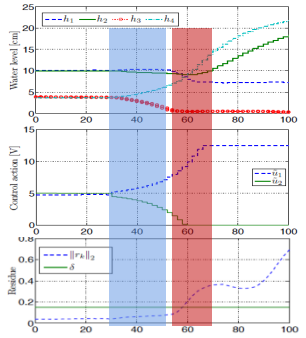
- Example: connect tank 3 to tank 1:

$$\Delta A = \begin{bmatrix} 0 & 0 & 0.0397 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -0.0402 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$





Summary



- Zero-dynamics attacks are robust to initial condition mismatch

- Proposed methods to reveal attacks by
  - Changing C: Adding measurements
  - Changing A: Modifying the open-loop dynamics
  - Changing B: Cooperatively scaling the input signals

- Adding measurements and scaling input signals does not affect the system performance

ROYAL INSTITUTE OF TECHNOLOGY

Summary of Lecture 3

- Risk management
  - security metrics, quantifying impact [1]
  - Tools for quantitative trade-off analysis between attacker's impact and resources: Important for defense prioritization
- Basics of Fault Diagnosis (for detectable attacks) [2]
- Watermarking (against undetectable attacks)
  - Induce asymmetries in the attacker/system models
  - Additive signals against sensor replay attacks [3]
  - Model perturbations against zero dynamics attacks [4]