



SINTEF

# GoHydro

Hans Ivar Skjelbred

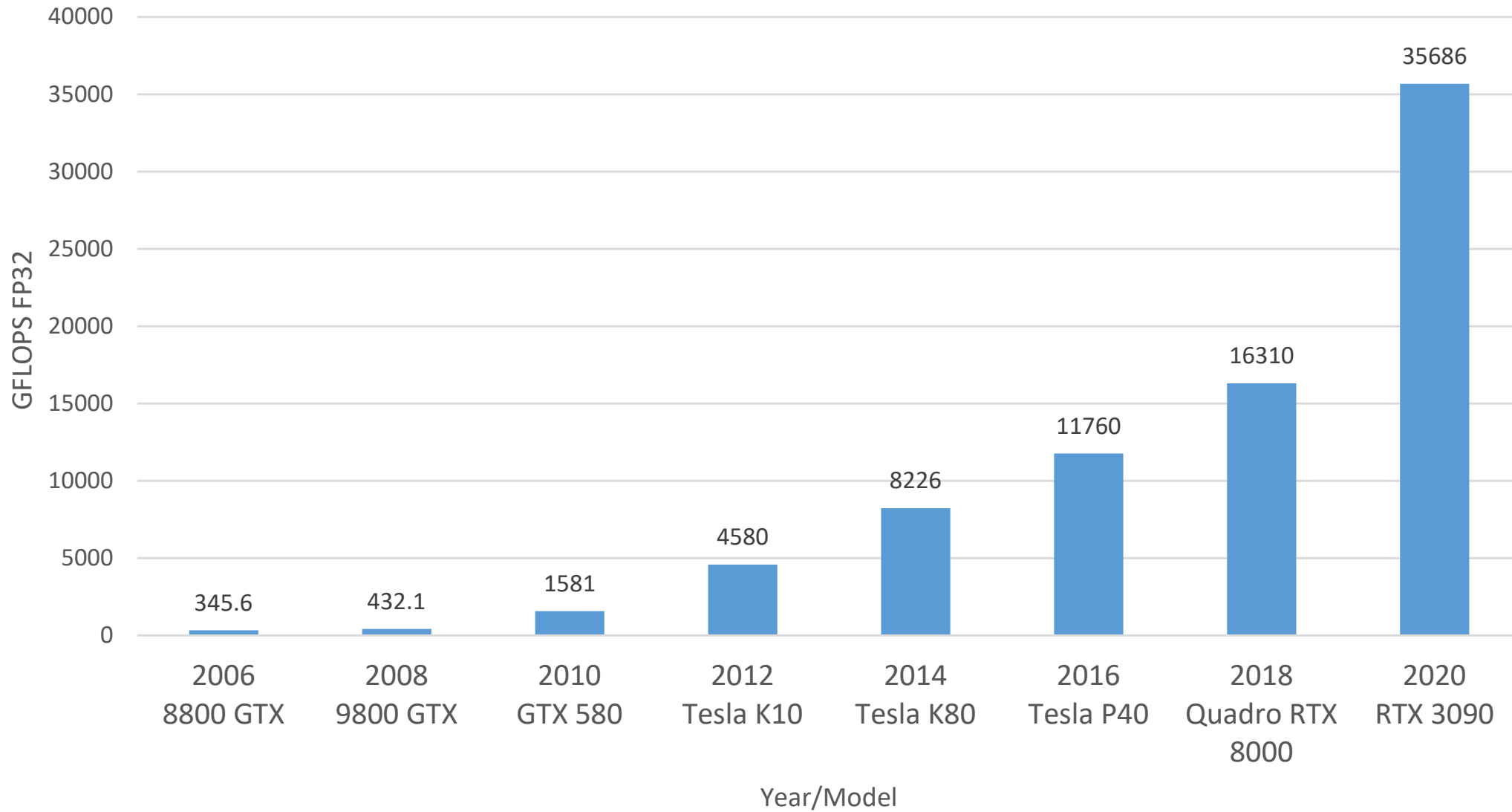
User Meeting – 17.11.2021



SINTEF

- Non-linear optimization problem
- GPU speed vs CPU speed
- Scaling properties without algorithmic details

# GPU performance development



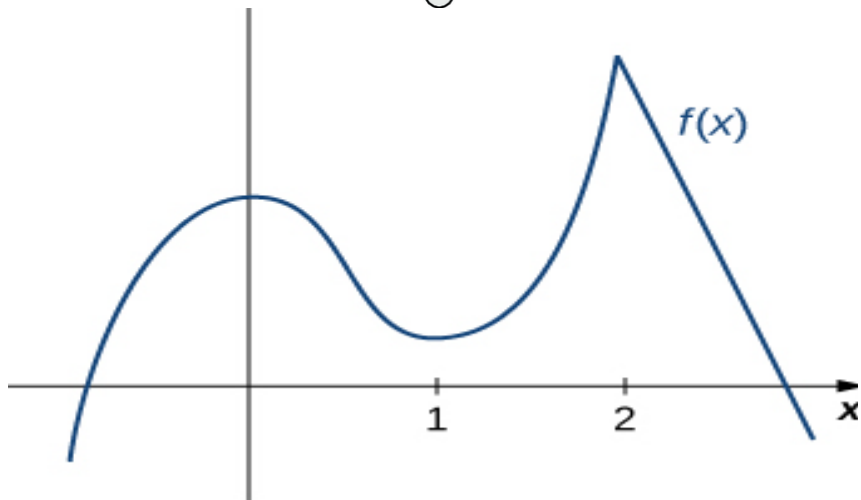
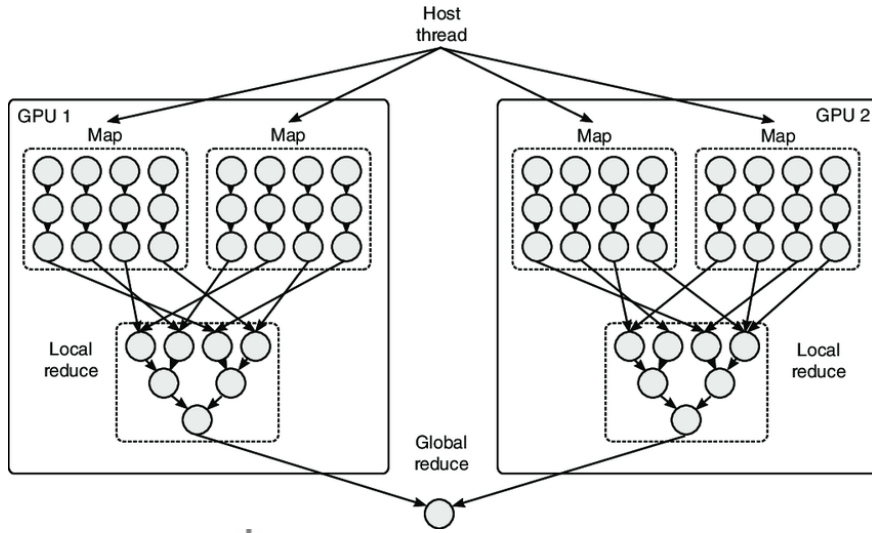
# KS Digitalisering 2020

Dele kunnskap og bygge opp infrastruktur for mer utstrakt bruk av GPU-akselerering i SINTEF

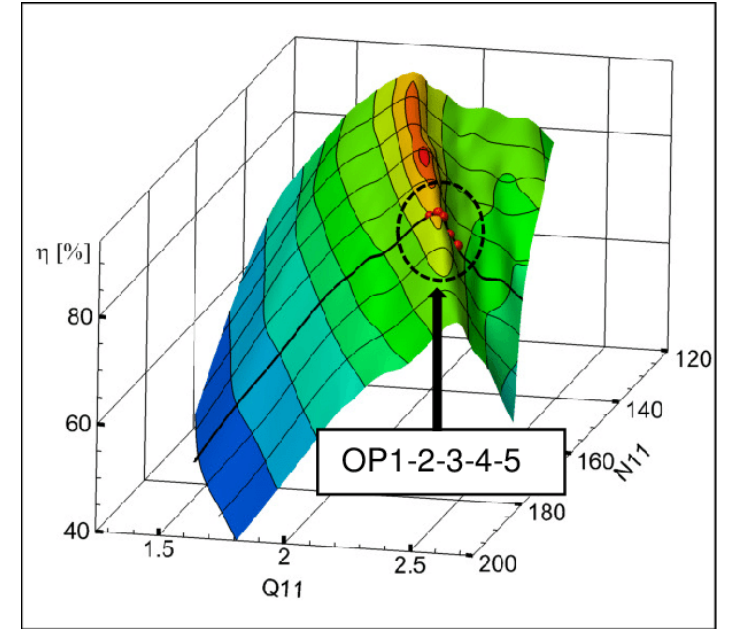
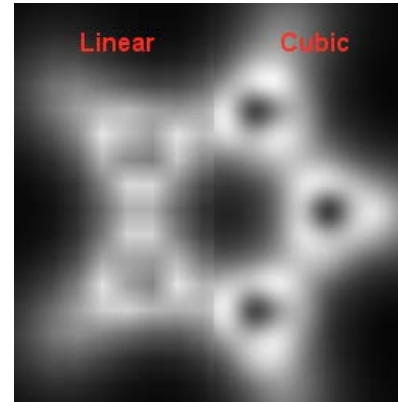


# GPU functions

- Maximization



- Interpolation



- Sum product

$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} + \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

HMMA FP16 or FP32  
 IMMA INT32

FP16  
 INT8 or UINT8

FP16  
 INT8 or UINT8

FP16 or FP32  
 INT32

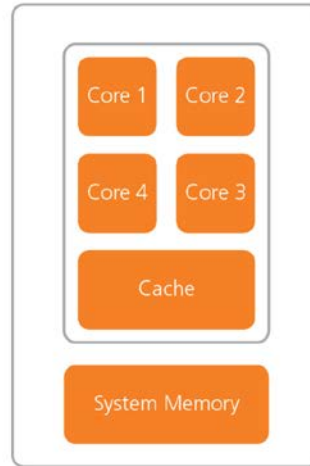


SINTEF

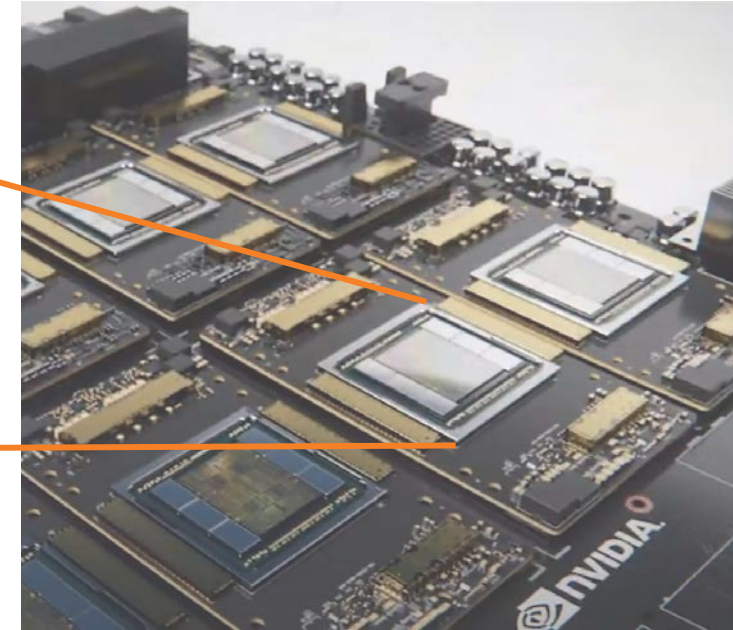
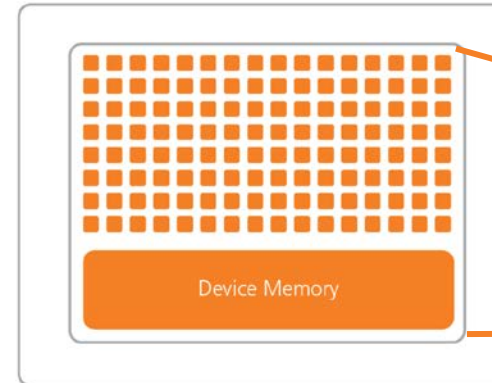
# GoHydro

- >100 000 identical tasks that can be solved in parallel
- Efficient memory addressing

CPU (Multiple Cores)



GPU (Hundreds of Cores)



Parallellize search for optimum

Maximize parallellization of the scheduling problem

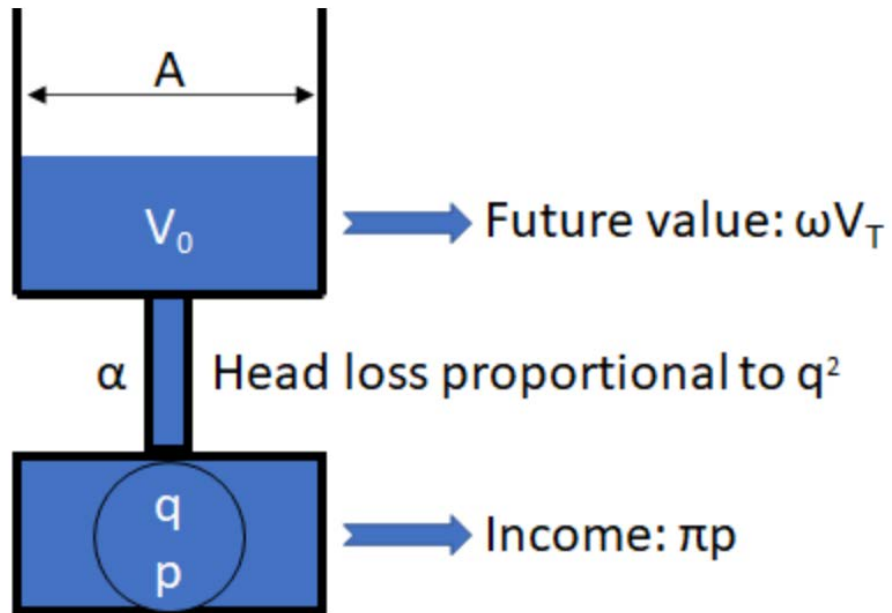
Scalable algorithm for non-linearities and large-scale problems



SINTEF

- Selection of essential modelling capabilities

# Verification vs analytical optimum of non-linear scheduling problem



Objektivfunksjon

$$\max \pi_1 p_1 + \pi_2 p_2 + \omega v_{end}$$

Betingelser

$$p_1 = Gq_1 \left( \frac{V_0 + V_0 - \gamma q_1}{2A} - \alpha q_1^2 \right)$$

$$p_2 = Gq_2 \left( \frac{V_0 - \gamma q_1 + V_0 - \gamma q_1 - \gamma q_2}{2A} - \alpha q_2^2 \right)$$

$$v_{end} = V_0 - \gamma q_1 - \gamma q_2$$

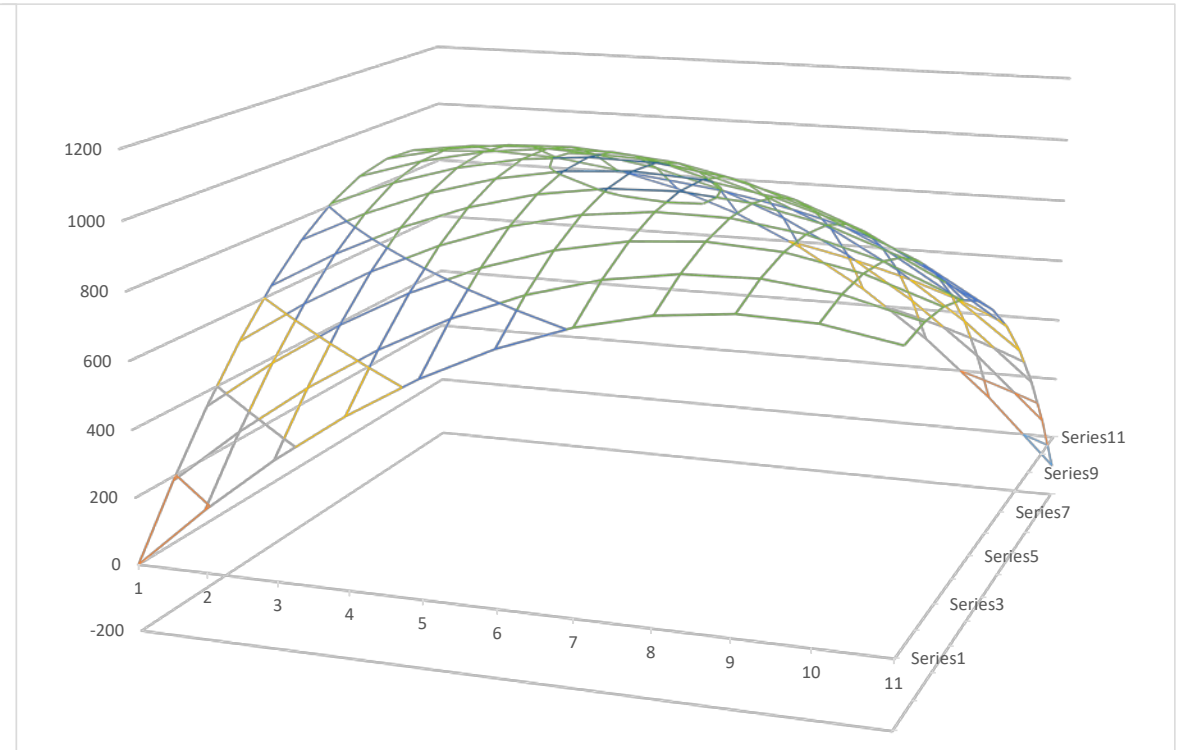
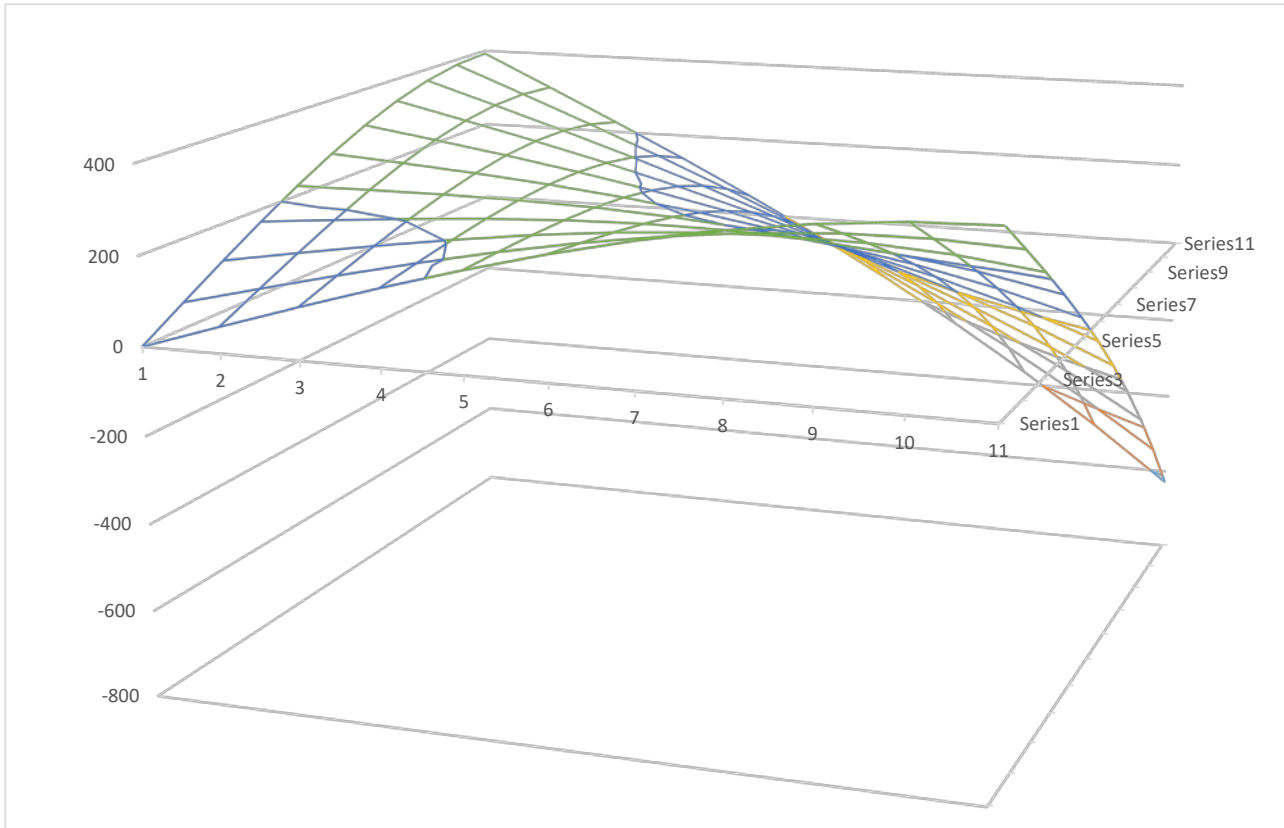




SINTEF

# Convexity for non-linear relations

- Compare head optimization with current tools





SINTEF

# Current work on algorithmic properties

- Upper bound from decomposed problem
- Lower bound from primal heuristic solution generation
- CPU for investigating iteration logic
- GPU for testing parallelization on various problem sizes
  - memory bandwidth
  - CUDA core computation occupancy

	scen1_prod	scen2_prod	scen3_prod	scen1_prod2	scen2_prod2	scen3_prod2
i1	0	28.21983	75.000679			
i2	0	33.617931	0			
i3	0	0	0	41.437439	55.149017	68.449059
SHOP/SHARM	0	0	0	41.44	55.25	68.45



SINTEF

Teknologi for et  
bedre samfunn