ASKAP Data Processing Using Apache Flink Streaming Platform

Muhammad Saad

Department of Computer Science(IFI)

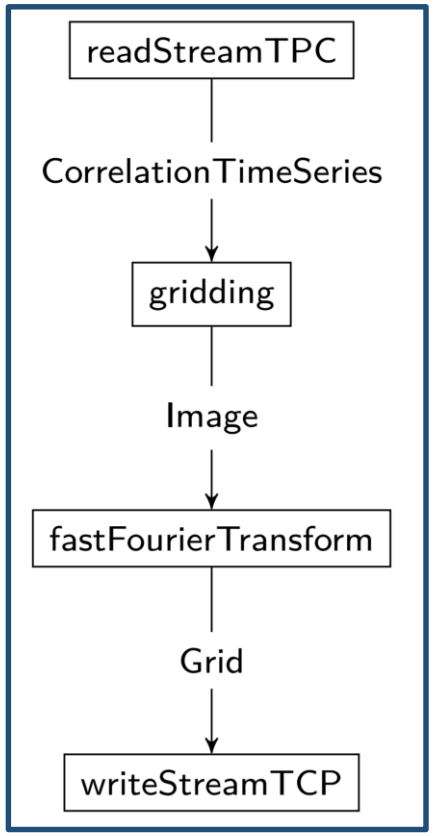
University of Zurich

Swiss SKA Days 2018

12 June, 2018

Privacy Preserving, Peta-scale Stream Analytics for Domain-Experts

- **Goal of project:**
  - **Data**
    - High Velocity
    - High Volume
  - Perform **Real Time Processing or continuously analyze** incoming data without storing it.(Streaming Processing)
- **ASKAP**(Australian Square Kilometer Array PathFinder) Data
  - 36 12 meter-antennas
  - Every dish generates 18 TB of samples per second
  - Gets reduced to 0.6 TB/s by beam formers
  - Then correlator computes correlations between the samples reducing it to a total of **2.5 GB/s** of raw data.



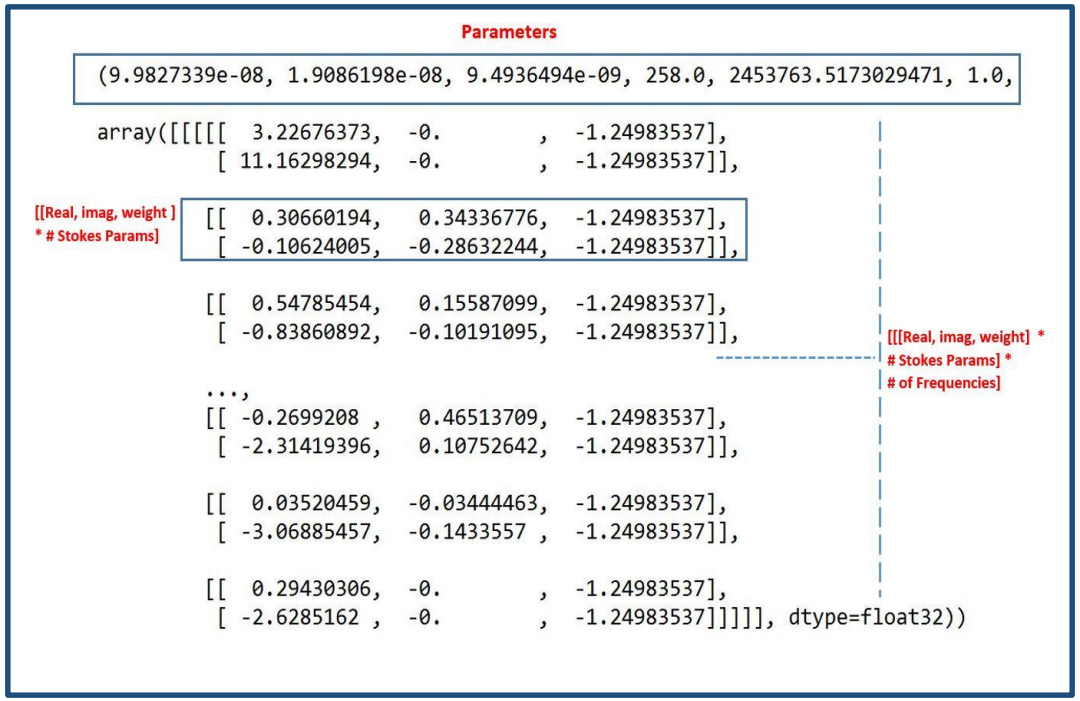
ASKAP Data Format

- **FITS / Measurement Sets**
- Each uv point consists of complex visibilities observed at multiple frequencies.
  - u_1,v_1
    - $V_1(u_1,v_1)$, $V_2(u_1,v_1)$, $V_3(u_1,v_1)$,..... $V_{nf}(u_1,v_1)$
  - u_2,v_2
    - $V_1(u_2,v_2)$, $V_2(u_2,v_2)$, $V_3(u_2,v_2)$,..... $V_{nf}(u_2,v_2)$

    .

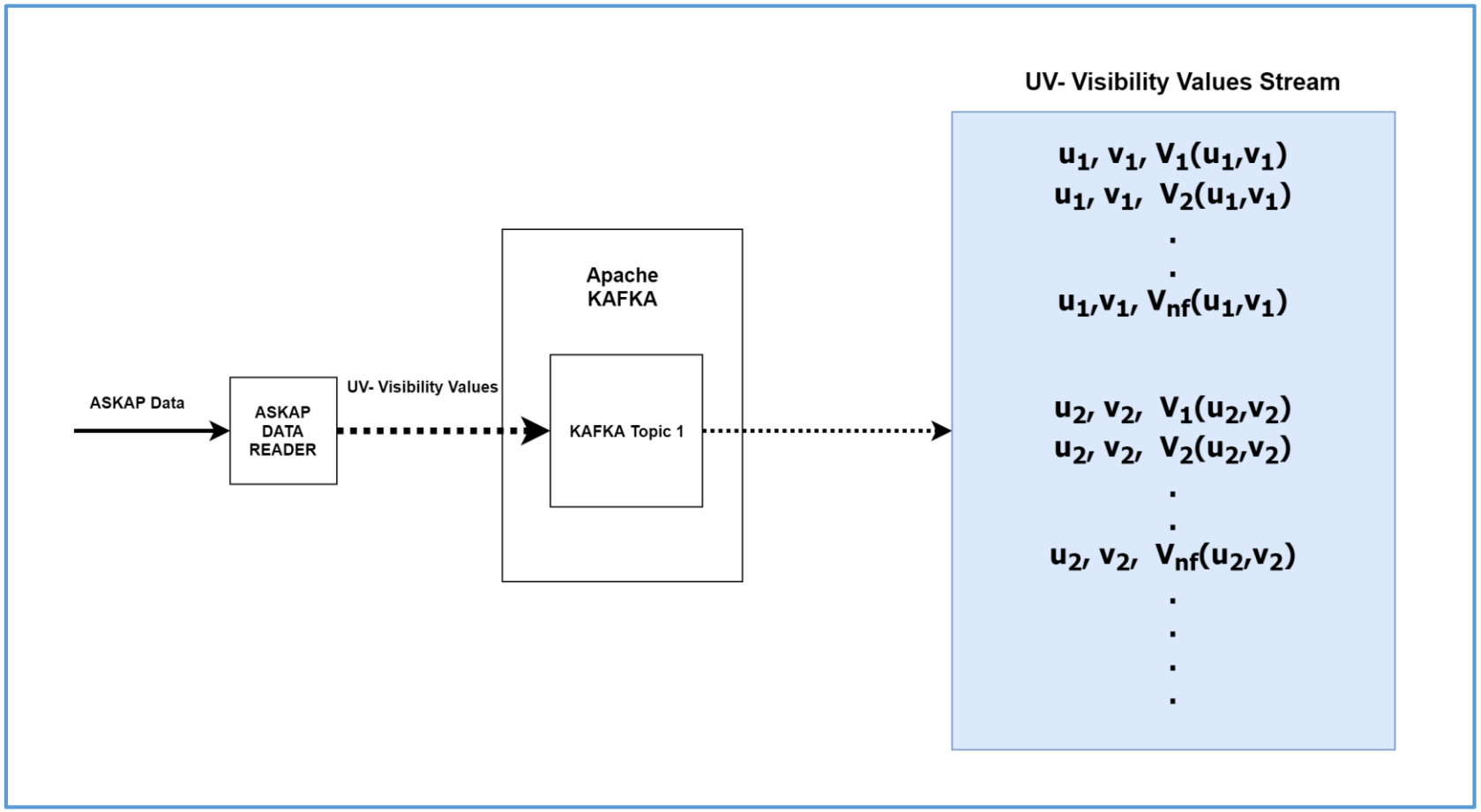
    .

    .
  - u_n,v_n
    - $V_n(u_1,v_1)$, $V_n(u_1,v_1)$, $V_n(u_1,v_1)$,..... $V_{nf}(u_n,v_n)$



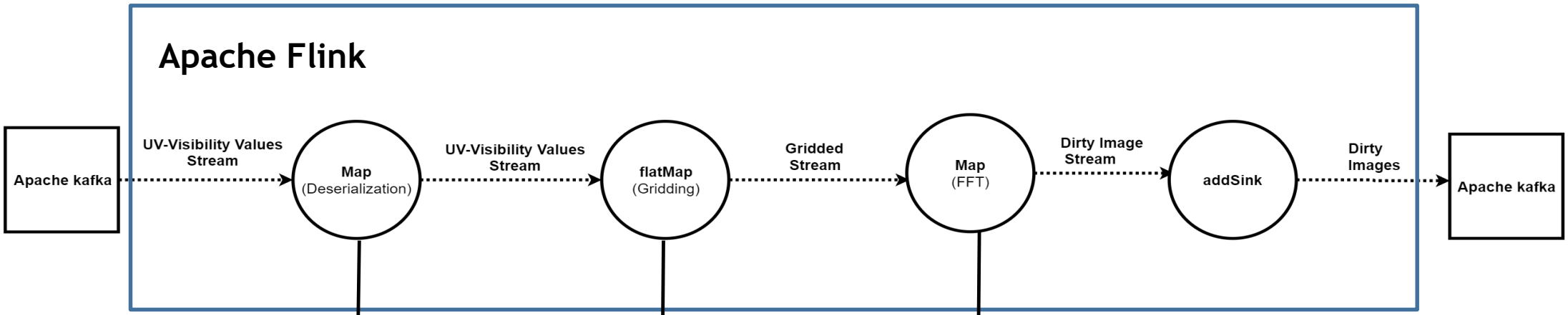
ASKAP Data Stream

- **Apache Kafka:** Stores Stream of data safely in a replicated fault tolerant manner.
- Visibility values vector grained finely to **UV Visibility Value pair**.



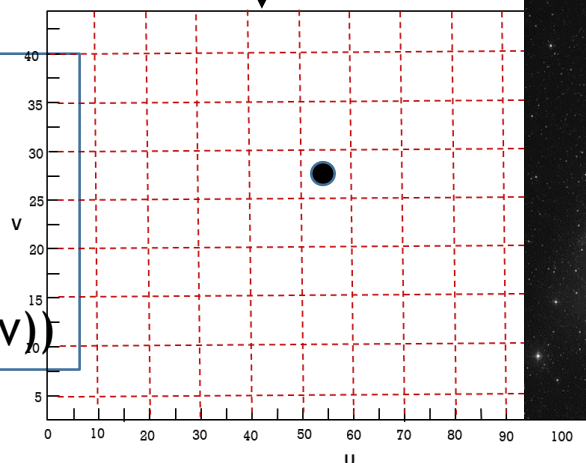
ASKAP Pipeline using Apache Flink

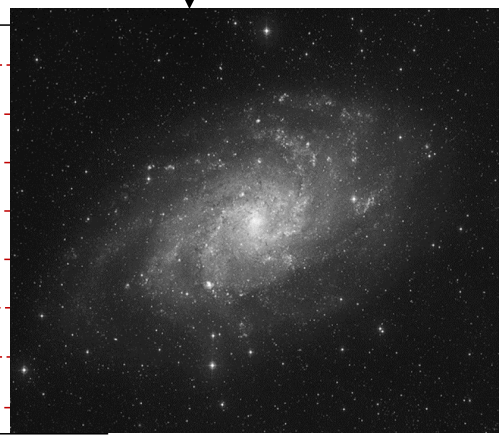
- Apache Flink is a scalable Stream Processing platform.



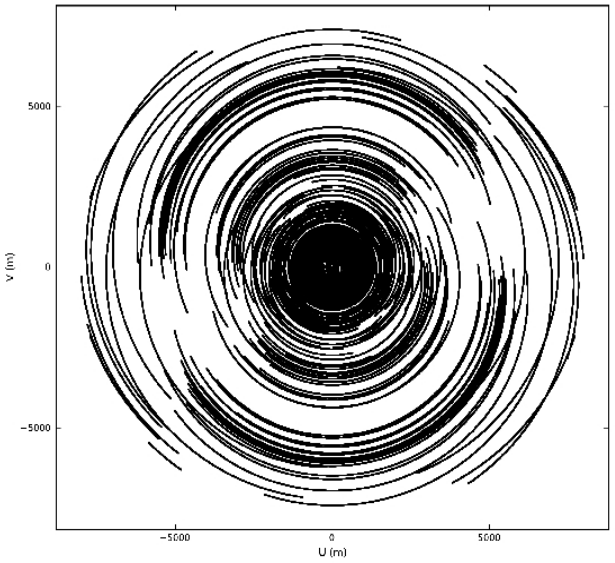
String(u,v,V(u,v))

Float(u,v)
Complex Float(V(u,v))

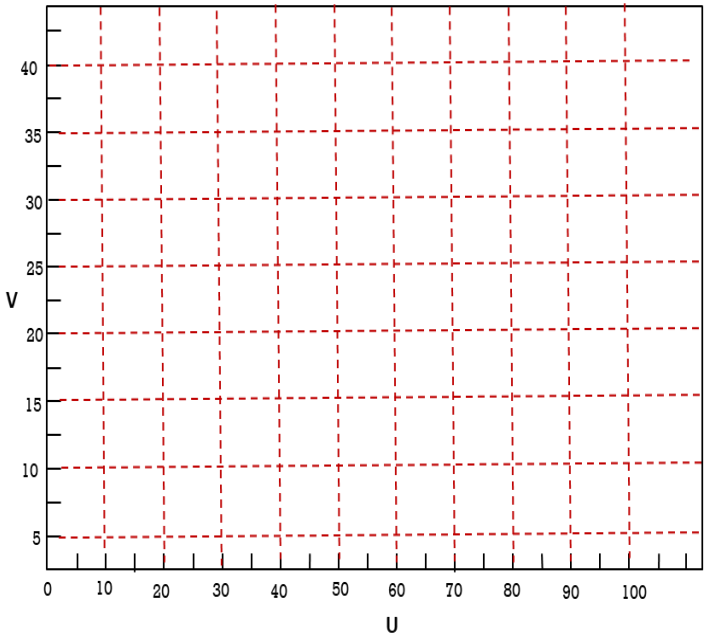




Real Time Processing?



UV- Coverage



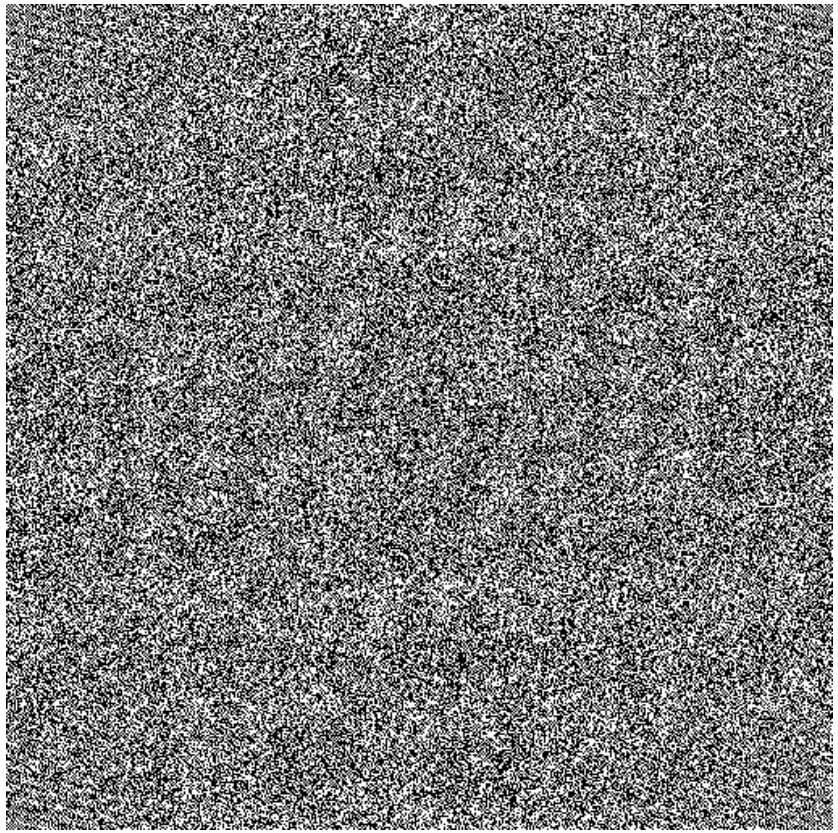
Grid

Dirty Image

Test Image from Fourier to Spatial Domain

Fourier Domain

512



512

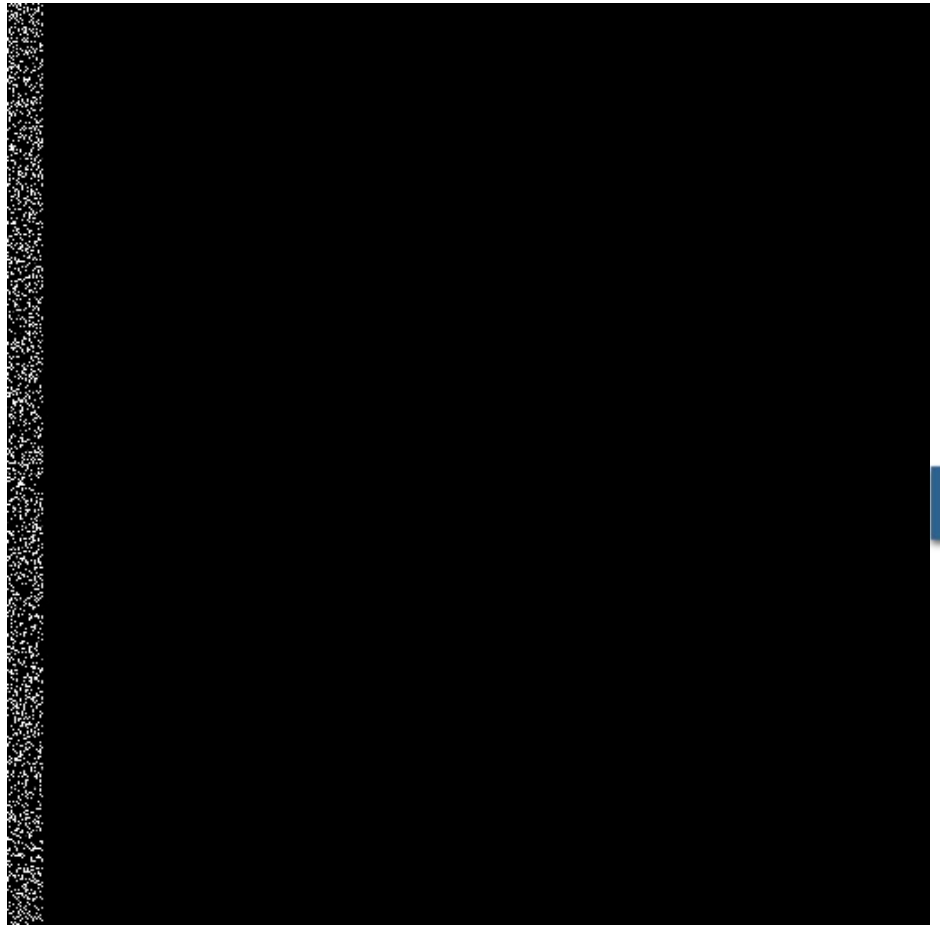
Spatial Domain

512

512

5k Points

Fourier Domain

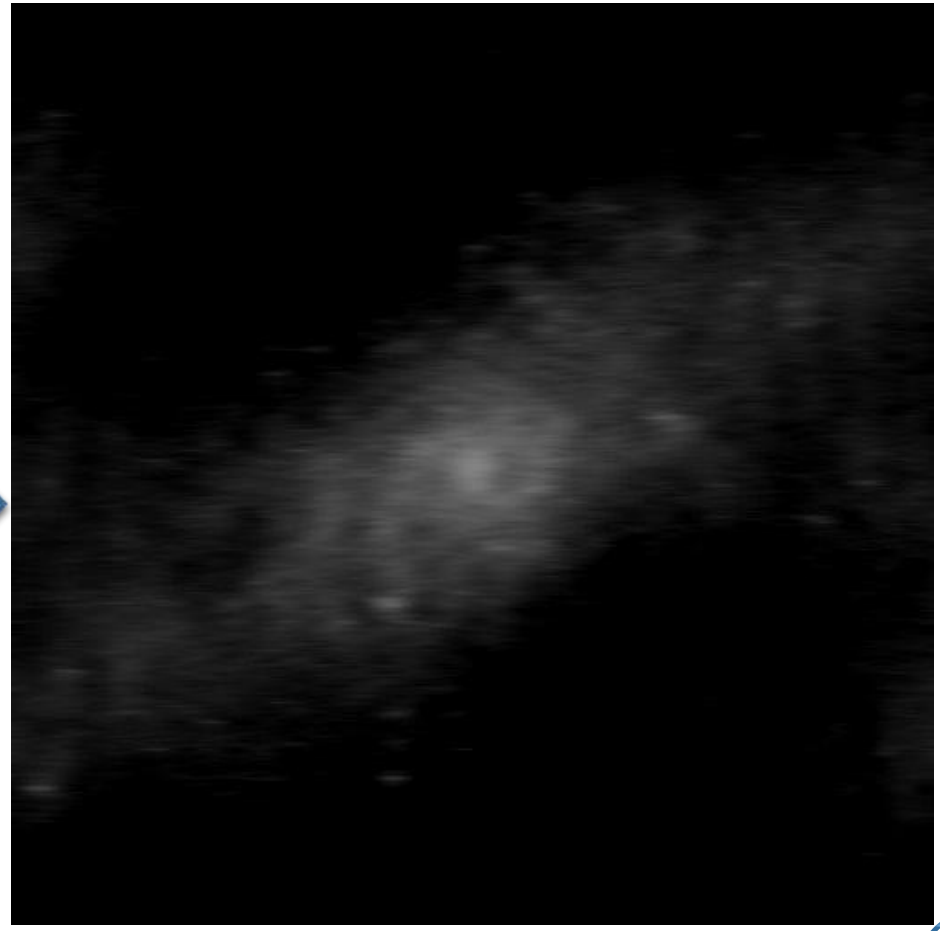


512

512

Inverse FFT

Spatial Domain

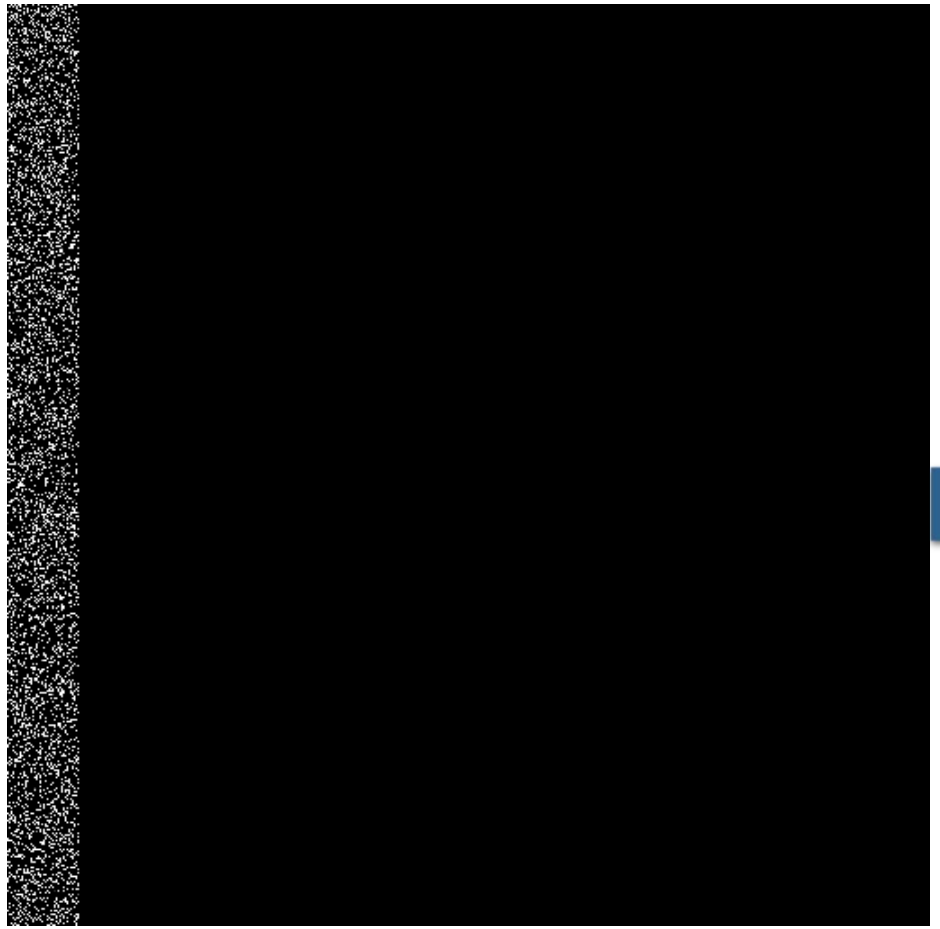


512

512

10k Points

Fourier Domain

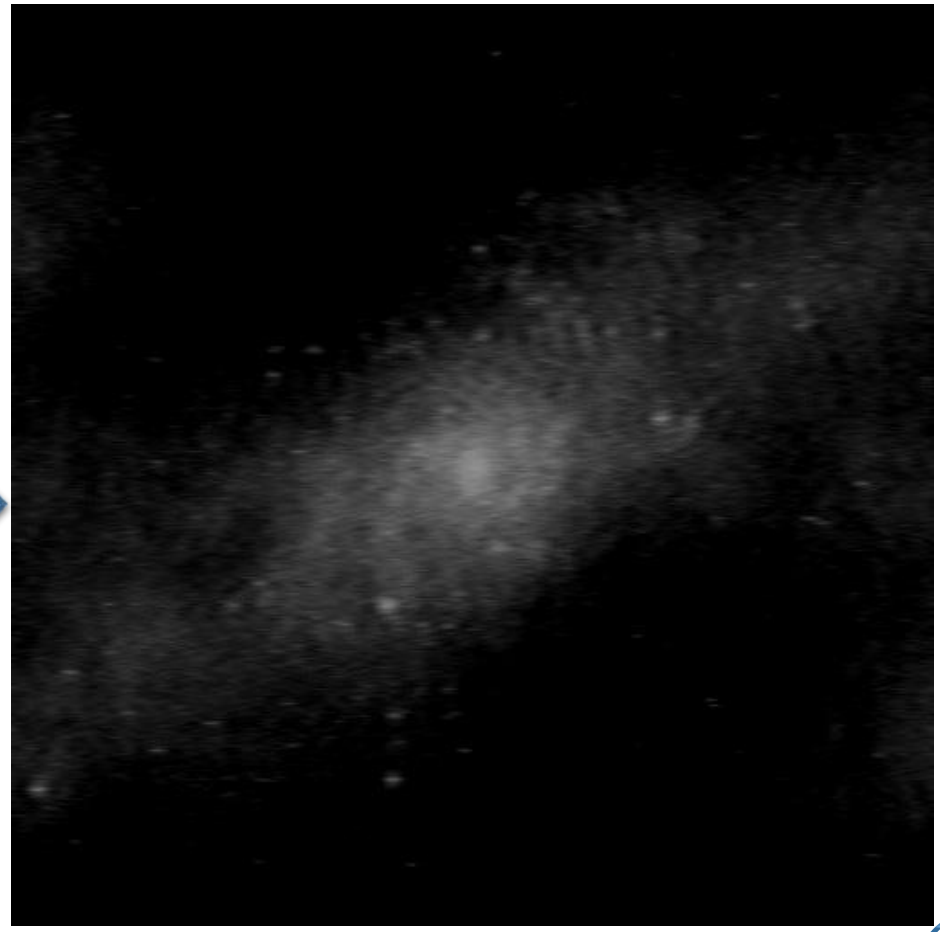


512

512

Inverse FFT

Spatial Domain

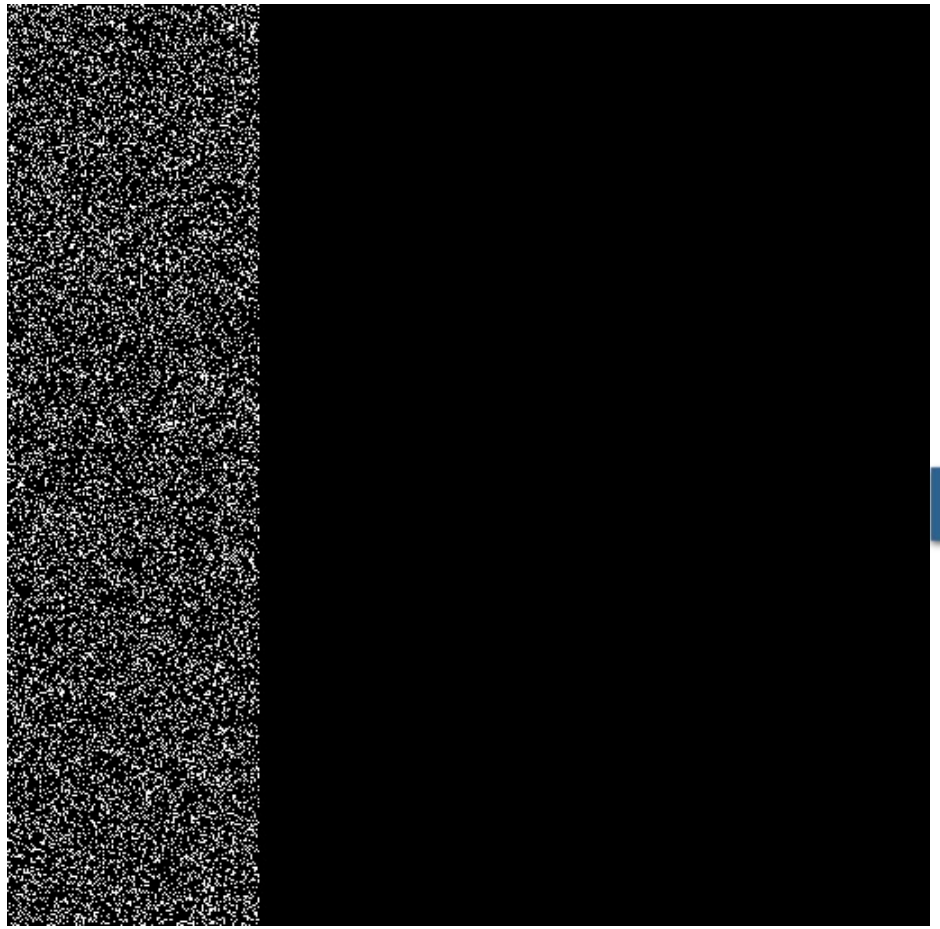


512

512

15k Points

Fourier Domain

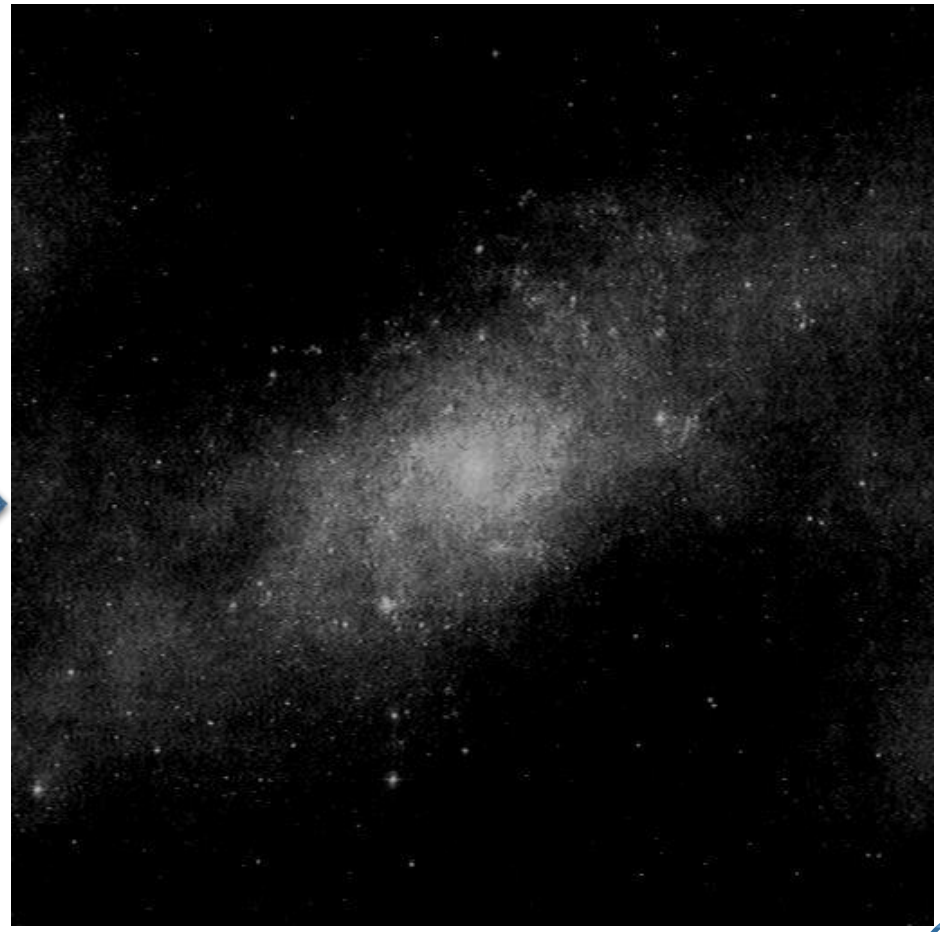


512

512

Inverse FFT

Spatial Domain

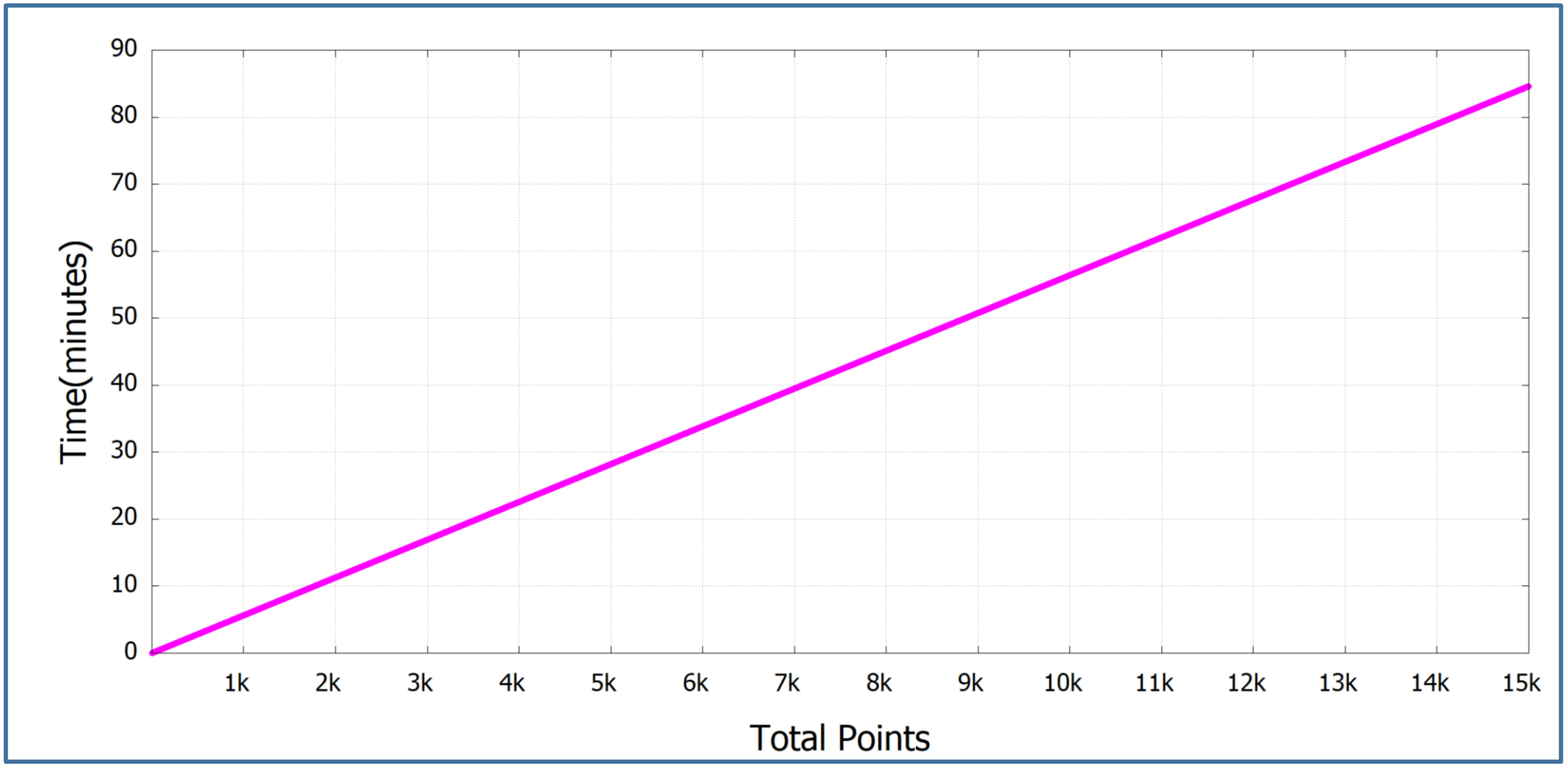


512

512

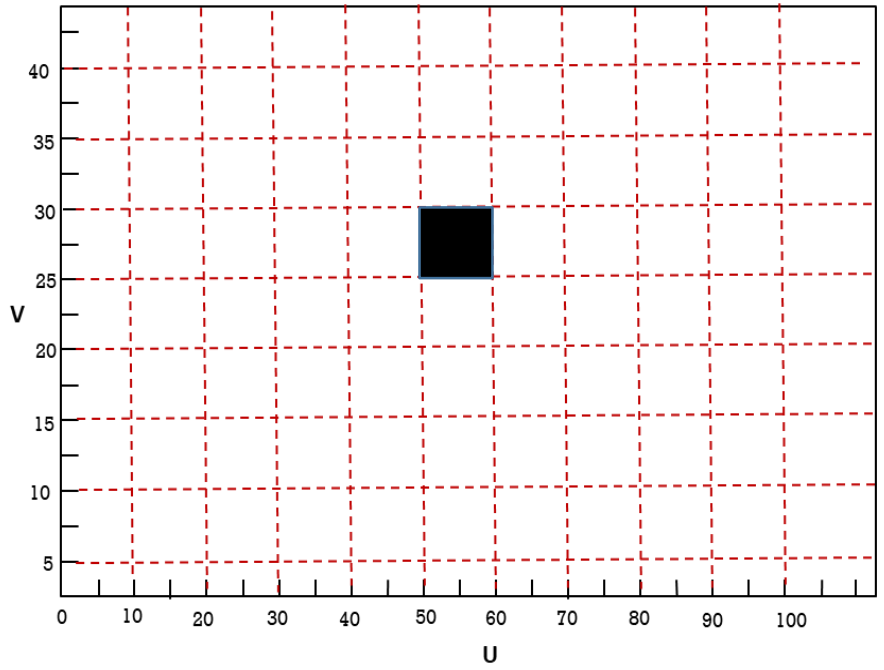
Cumulative Runtime for FFT

- Cumulative Time
  - By iteratively adding one point to grid
  - Then FFT on the Grid (512 x 512)



Incremental FFT

- Idea:
  - If any point is added in the grid
  - Then Fourier Transform calculations should bound to that point only
  - Not every point already present in the grid



Incremental FFT

$$F(x,y) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f(m,n) e^{(-i2\pi)(x\frac{m}{M} + y\frac{n}{N})}$$

x,m = 0,1 … M-1
y,n = 0,1 … N-1

- For any point f(i,j) and consider $\omega = e^{-2\pi i}$

F(x,y)=

$$\begin{bmatrix}
f(i,j)\cdot\omega^{(0)(\frac{i}{M})}\cdot\omega^{(0)(\frac{j}{N})} & f(i,j)\cdot\omega^{(0)(\frac{i}{M})}\cdot\omega^{(1)(\frac{j}{N})} & f(i,j)\,\omega^{(0)(\frac{i}{M})}\cdot\omega^{(2)(\frac{j}{N})} & \dots & f(i,j)\,\omega^{(0)(\frac{i}{M})}\cdot\omega^{(N-1)(\frac{j}{N})} \\
f(i,j)\cdot\omega^{(1)(\frac{i}{M})}\cdot\omega^{(0)(\frac{j}{N})} & f(i,j)\cdot\omega^{(1)(\frac{i}{M})}\cdot\omega^{(1)(\frac{j}{N})} & f(i,j)\cdot\omega^{(1)(\frac{i}{M})}\cdot\omega^{(2)(\frac{j}{N})} & \dots & f(i,j)\cdot\omega^{(1)(\frac{i}{M})}\cdot\omega^{(N-1)(\frac{j}{N})} \\
f(i,j)\cdot\omega^{(2)(\frac{i}{M})}\cdot\omega^{(0)(\frac{j}{N})} & f(i,j)\cdot\omega^{(2)(\frac{i}{M})}\cdot\omega^{(1)(\frac{j}{N})} & f(i,j)\cdot\omega^{(2)(\frac{i}{M})}\cdot\omega^{(2)(\frac{j}{N})} & \dots & f(i,j)\cdot\omega^{(2)(\frac{i}{M})}\cdot\omega^{(N-1)(\frac{j}{N})} \\
 & & \vdots & & \\
f(i,j)\cdot\omega^{(M-1)(\frac{i}{M})}\cdot\omega^{(0)(\frac{j}{N})} & f(i,j)\cdot\omega^{(M-1)(\frac{i}{M})}\cdot\omega^{(1)(\frac{j}{N})} & f(i,j)\cdot\omega^{(M-1)(\frac{i}{M})}\cdot\omega^{(2)(\frac{j}{N})} & \dots & f(i,j)\cdot\omega^{(M-1)(\frac{i}{M})}\cdot\omega^{(N-1)(\frac{j}{N})}
\end{bmatrix}$$

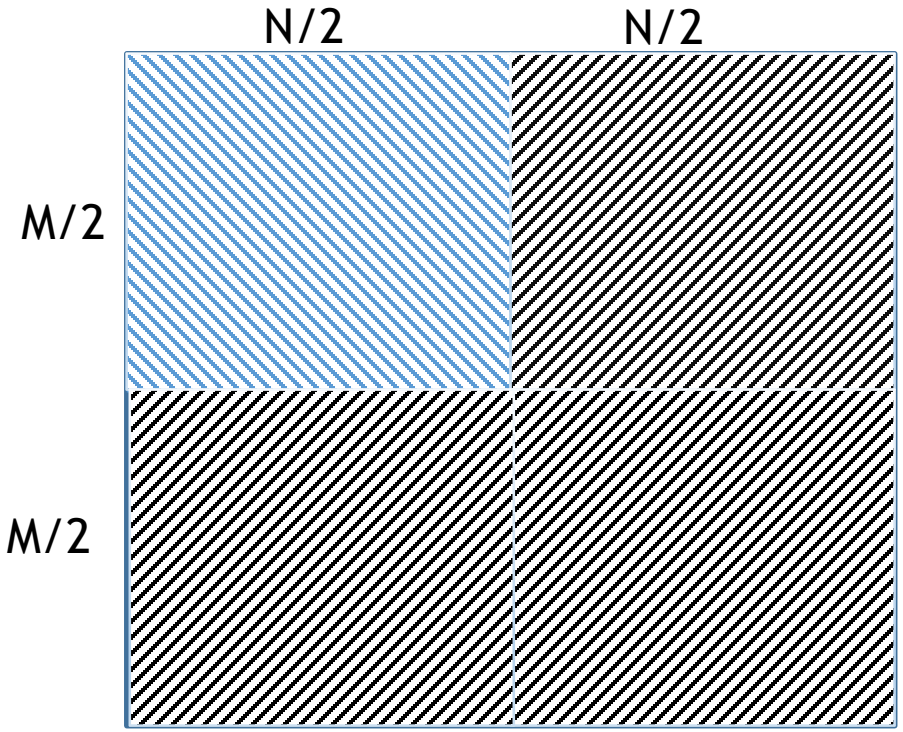
Incremental FFT

- By symmetry:

$$\text{if } \omega_N^k = e^{\frac{-2\pi ik}{N}}; \quad \omega_N^{k+\frac{N}{2}} = -\omega_N^k$$

- For any point f(i,j)

F'(x,y) =

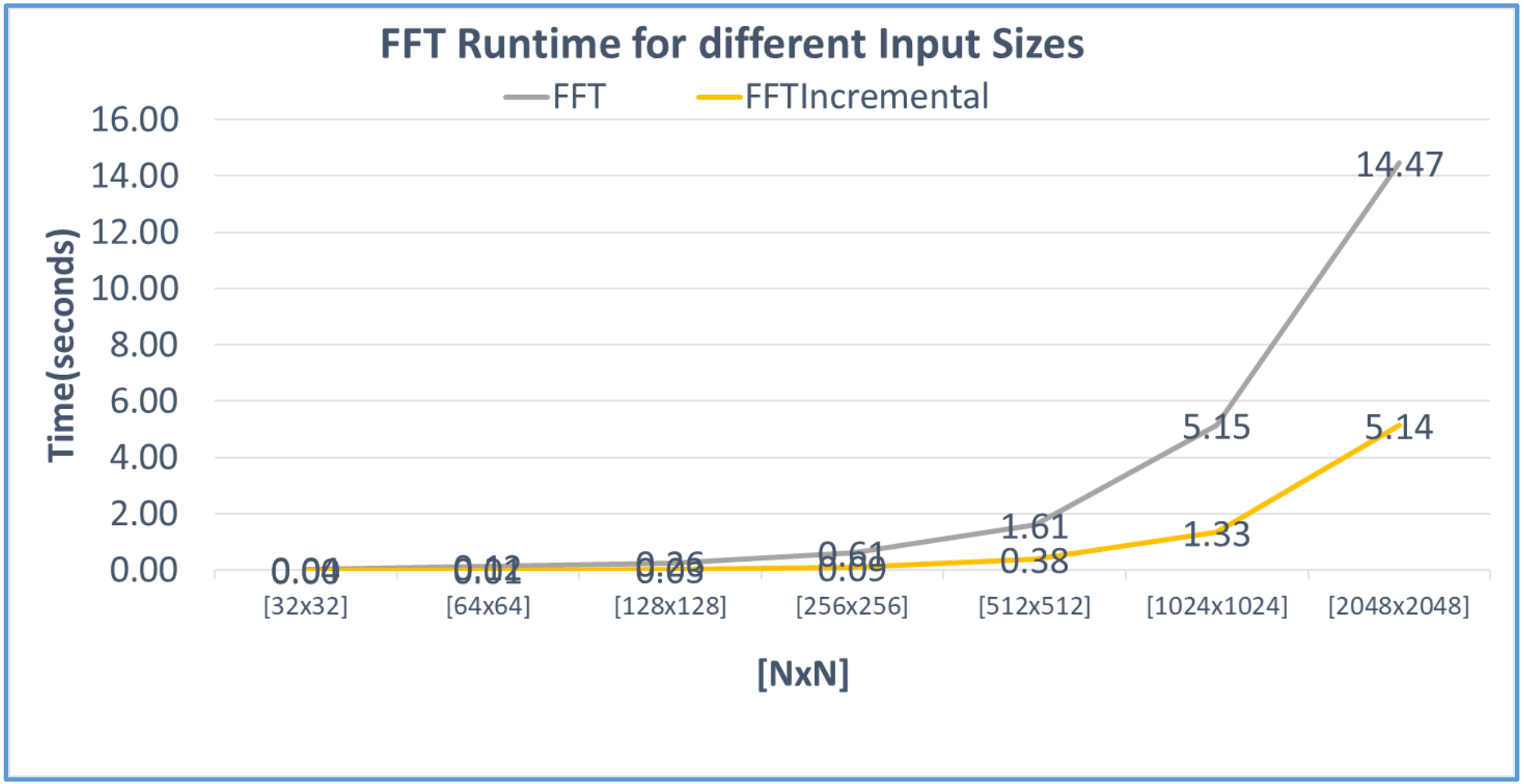


Add F'(x,y) calculated for one point to Fourier Matrix

F(x,y) = F(x,y) + F'(x,y)

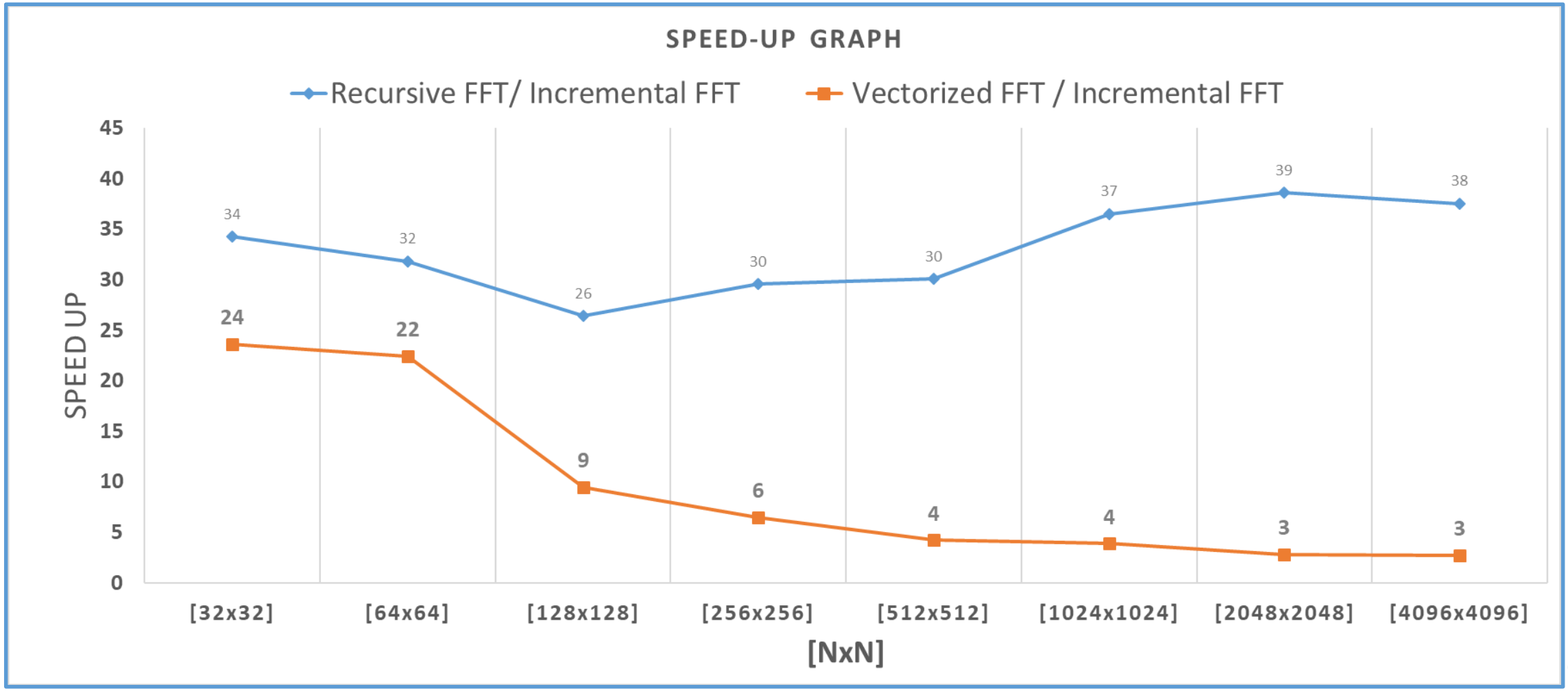
FFT vs Incremental FFT Runtime

- Runtime by adding 1 point in Grid and then taking FFT



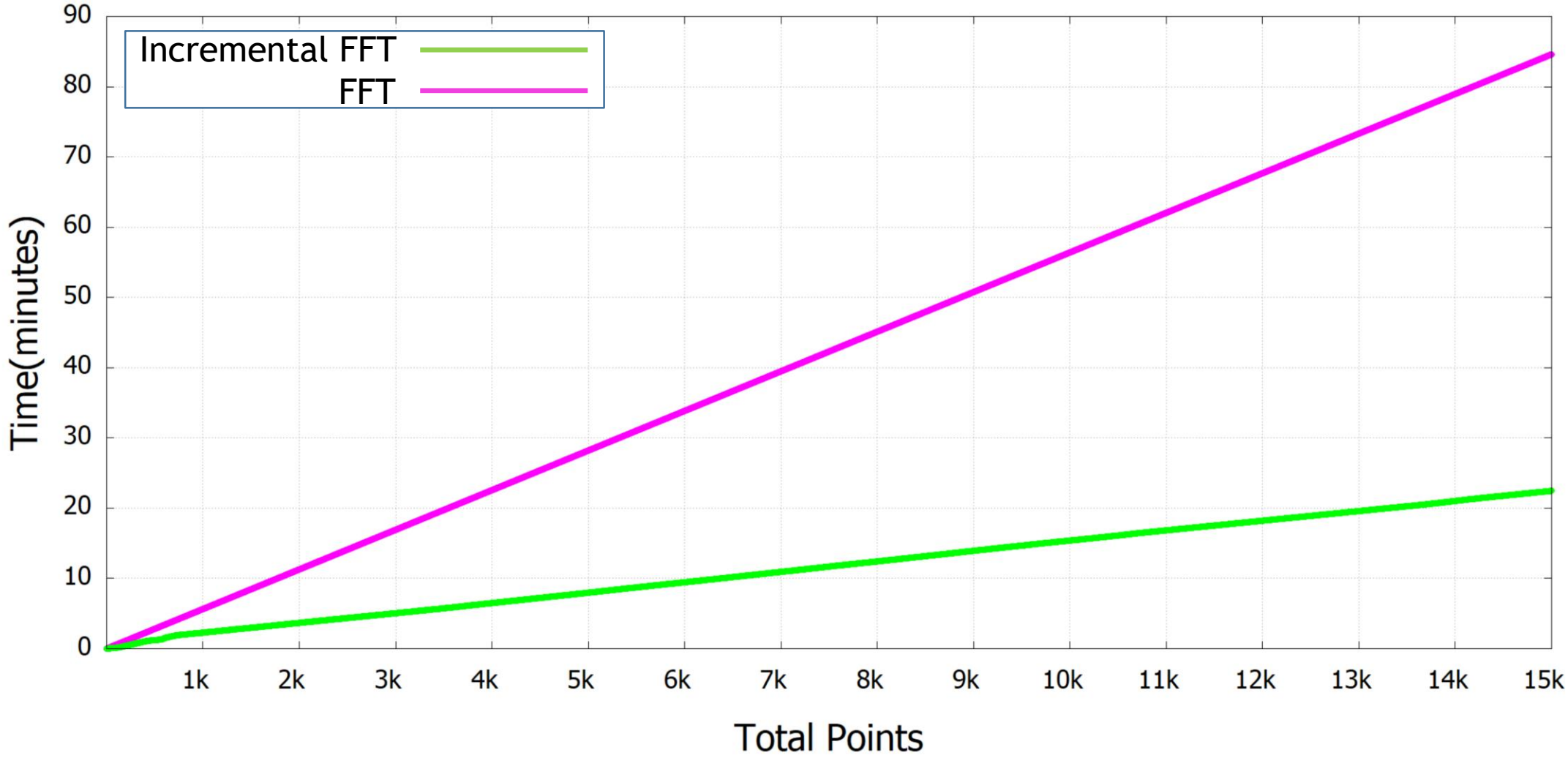
Incremental FFT speedup over FFT

- Speedup by adding 1 point in Grid and then taking FFT



Cumulative Runtime for FFT and Incremental FFT

Speedup for only 1 incoming point in a stream



Future Work

- Optimization and improvement of single point Incremental FFT algorithm as well as implementation
- Incremental FFT over **multiple** points rather than **single** point.
  - Specify Time Window
  - Specify number of points
- Complete the operations in the Streaming pipeline(Weighting)

Thanks … !

Questions