Caches as a new PHY Resource in 6G?

(WFF 2023)

Petros ELIA

(Professor - Communication Systems Department)
EURECOM
Sophia Antipolis, France
Video on-demand (VoD): ≥ 70% of traffic

Communicating VoD is very expensive

For wireless network providers
- Costly bandwidth and infrastructure

Content providers (Netflix, Amazon, etc.) pay large fees per content
Pros for Caches in PHY

➢ Most VoD content is cacheable
➢ Networks are already full of caches (your cell phone)
➢ Caches are currently used the wrong way (push closer to users)

Consider Caches as a New Resource for PHY

➢ Use caches to alter structure (not volume) of the PHY problem
Historically a new resource (e.g. MIMO) brings:

- Big algorithmic gains
- Interesting algorithmic challenges
- New resource must work well with existing solutions
New resource applies in several downlink scenarios

**Wired or wireless communications**

**Satellite communications**

**In-flight communications**

**Virtual Reality**

**Cloud computing**
Let’s slow down: Original idea of Coded Caching

Key breakthrough: USE CACHES TO CANCEL INTERFERENCE

Result: Maddah-Ali, Niesen (Bell Labs - 2013)
Example:  \( N = K = 3, M = 2 \)  \( (\gamma = \frac{2}{3}) \)

Library: \( N = 3 \) files

\[
\begin{array}{ccc}
A_{12} & A_{13} & A_{23} \\
B_{12} & B_{13} & B_{23} \\
C_{12} & C_{13} & C_{23}
\end{array}
\]

\[
\begin{array}{c}
\text{Rx1} \\
\text{Rx2} \\
\text{Rx3}
\end{array}
\]

\[
\begin{array}{c}
M = 2 \\
M = 2 \\
M = 2
\end{array}
\]
Example: $N = K = 3, M = 2$ \quad \left( \gamma = \frac{M}{N} = \frac{2}{3} \right)$

Library: $N$ files

$A_{12}$  $A_{13}$  $A_{23}$

$B_{12}$  $B_{13}$  $B_{23}$

$C_{12}$  $C_{13}$  $C_{23}$

$\frac{1}{3}$
Example: \( N = K = 3, M = 2 \) \( (\gamma = \frac{M}{N} = \frac{2}{3}) \)

- **Library:** \( N \) files

- **Transmit:** \( A_{23} \oplus B_{13} \oplus C_{12} \) (a common message for all)

- **Speedup Factor** = 3 = \( Ky + 1 \) users at a time
Can caching truly make a difference as a core technology in wireless communications?

\[ K = 3, \gamma = \frac{2}{3}, K\gamma = 2 \]

Coded Caching: Intuition - Clique
Can caching truly make a difference as a core technology in wireless communications?

Coded Caching: Intuition - Cliques

\[ K = 4, \quad \gamma = \frac{2}{4}, \quad K\gamma = 2 \]

\[
\begin{array}{c}
K = 4, \quad \gamma = \frac{2}{4}, \quad K\gamma = 2
\end{array}
\]
Coded Caching: Intuition – Problematically Many Cliques

\[ K = 100, \quad \gamma = \frac{9}{100} \quad K\gamma = 9 \]

10 users at a time, \( 2 \cdot 10^{13} \) cliques
Bad news: There exist ‘fundamental’ bottlenecks

• In Theory: Speedup = $K\gamma + 1$

• *In theory, theory and practice are the same. In practice they are not.*” Albert Einstein
Bad news: There exist ‘fundamental’ bottlenecks
Resolve fundamental limitations.

• Coded Caching requires infinite SNR
  • Gain goes to zero for smaller SNR
  • Reason: Worst-user multicasting

• Requires ASTRONOMICAL file sizes
  • Gain very small for reasonable file sizes

• Works for one transmit antenna
  • Discard multi-antenna systems? Absurd.
Elevating to multi-antennas: Resolving file-size problem

- Gains vanish if files are not astronomically large!

Because we need astronomically many cliques

If you don’t have astronomical file sizes, the gains almost VANISH

Effective (MAX) \[ K \gamma + 1 \approx 5 - 7 \]

Simple Solution: two birds with one stone

Massive CACHING gains
Can caching truly make a difference as a core technology in wireless communications?

Resolving file-size problem (Lampiris-Elia JSAC)

$L$ transmit antennas

Able to multiplicatively boost multiplexing gains

Subpacketization

\[ S_{new} < \sqrt[\large L]{S_{old}} \]

e.g. $L = 5$

$10^8$GBytes $\rightarrow$ 120 Bytes
**Massive boost of speedup factor**

<table>
<thead>
<tr>
<th>Harsh Reality</th>
<th>Limited File Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adding Antennas</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{align*}
1 + 7 & \rightarrow 8 \\
2 + 7 & \rightarrow 16 \\
3 + 7 & \rightarrow 24 \\
4 + 7 & \rightarrow 32 \\
\vdots
\end{align*}
\]
New algorithm: Pyramid of vectors

\[ \mathbf{x}_{1234} = \mathcal{H}_{g_1}^{-1} \begin{bmatrix} W_{1234}^1 \\
W_{1234}^2 \\
W_{2134}^3 \\
W_{2134}^4 \\
\end{bmatrix} + \mathcal{H}_{g_2}^{-1} \begin{bmatrix} W_{134}^1 \\
W_{134}^2 \\
W_{134}^3 \\
W_{134}^4 \\
\end{bmatrix} + \mathcal{H}_{g_3}^{-1} \begin{bmatrix} W_{124}^1 \\
W_{124}^2 \\
W_{124}^3 \\
W_{124}^4 \\
\end{bmatrix} + \mathcal{H}_{g_4}^{-1} \begin{bmatrix} W_{124}^1 \\
W_{124}^2 \\
W_{124}^3 \\
W_{124}^4 \\
\end{bmatrix} \]

\[ \mathbf{y}_{1234}^{(1)} = \mathcal{H}_{g_1} \mathcal{H}_{g_1}^{-1} \begin{bmatrix} W_{1234}^1 \\
W_{1234}^2 \\
W_{2134}^3 \\
W_{2134}^4 \\
\end{bmatrix} + \mathcal{H}_{g_2} \mathcal{H}_{g_2}^{-1} \begin{bmatrix} W_{134}^1 \\
W_{134}^2 \\
W_{134}^3 \\
W_{134}^4 \\
\end{bmatrix} + \mathcal{H}_{g_3} \mathcal{H}_{g_3}^{-1} \begin{bmatrix} W_{124}^1 \\
W_{124}^2 \\
W_{124}^3 \\
W_{124}^4 \\
\end{bmatrix} + \mathcal{H}_{g_4} \mathcal{H}_{g_4}^{-1} \begin{bmatrix} W_{124}^1 \\
W_{124}^2 \\
W_{124}^3 \\
W_{124}^4 \\
\end{bmatrix} \]
Intuition: multiple decompositions with few antennas
This has been Theory

- Theoretical gains are large.
- What happens when we move closer to practice
- ERC Proof of Concept – LIGHT

Wired or wireless (Cellular-Wifi)

Satellite communications

Cloud computing
Focus on Multi-User MISO – Cellular, WiFi, etc.
Can caching truly make a difference as a core technology in wireless communications?

Gains over OPTIMIZED MISO BC – Analysis and Sims

- Optimize as you wish
- Account for CSI costs

\[
Gain = \frac{R^*_\Sigma (\text{cache} - \text{aided})}{R^*_\Sigma (\text{no} - \text{cache})}
\]

"Vector Coded Caching Multiplicatively Boosts the Throughput of Realistic Downlink Systems," Zhao-Bazco-Elia
Can caching truly make a difference as a core technology in wireless communications?

Large Matrix Analysis and Sims of Cache-Aided MIMO-BC

Gain = \frac{R_\Sigma^*(\text{cache – aided})}{R_\Sigma^*(\text{no – cache})}

"Vector Coded Caching Multiplicatively Boosts the Throughput of Realistic Downlink Systems," Zhao-Bazco-Elia
<table>
<thead>
<tr>
<th>$G$</th>
<th>Cache: 25GB</th>
<th>Movies: HD, 1.3GB, 90 minutes</th>
<th>Latency: 2 min. (small buffer OK)</th>
<th>Comm Packet: 50 bytes</th>
<th>Subpacketization: 600K</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G = 7$</td>
<td>Cache: 25GB</td>
<td>Movies: Full-HD, 2.47GB, 90 minutes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 6$</td>
<td>Cache: 25GB</td>
<td>Movies: SD-480p</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 5$</td>
<td>Cache: 5GB</td>
<td>Movies: SD-480p</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Netflix movies - Zipf $\approx 1.4$
90% of traffic speeds-up by theoretical factor of $G$
Can caching truly make a difference as a core technology in wireless communications?

Gains Across Various Precoders (MISO)

**Fig. 2:** Effective gain $G^*$ over optimized cacheless system for $L \in \{32, 64\}$ and $G = 6$. 
Most Challenging Scenario – very large L, Large Cells, Lower power

- 75% < -3dB
- 40% < -10dB
- 60% < 9dB
- 75% < 27dB
- 44% < 21dB
Can caching truly make a difference as a core technology in wireless communications?

Large cell – 32 Tx-Antennas

Macro-Cell:

\[ D_1 = 35 \]
\[ D_2 = 400 \]
\[ L = 32 \]
\[ M = 4 \]
\[ G = 6 \]

Optimal \( Q \) and \( Q' \)

![Graph showing effective coded caching gain vs. power level](chart)

- 75% < 9dB
- 50% < 9dB
Smaller cells – w. & w/o Max-Min Fairness MU-MIMO

Effective Coded Caching Gain

No User Fairness

Micro-Cell:
\[ D_1 = 10, D_2 = 100 \]
\[ \eta = 3, L_0 = 10^{-3}, r \]
\[ L = 32 \]
\[ M = 4 \]
\[ G = 6 \]
Optimal \( Q \) and \( Q' \)

Max-Min Fairness

Micro-Cell:
\[ D_1 = 10, D_2 = 100 \]
\[ \eta = 3, L_0 = 10^{-3}, r \]
\[ L = 32, M = 4 \]
\[ G = 6 \]
Optimal \( Q \) and \( Q' \)
Can caching truly make a difference as a core technology in wireless communications?

Many Applications to be Explored...

Wired or wireless communications

Satellite communications

In-flight communications

Virtual Reality

Cloud computing
Need and opportunity

✓ Abundant new resource
✓ Works well with other resources (very simple precoding)
✓ Very substantial gains

✓ Some VERY Simple algorithms
  ✓ E.g. Superimpose ZF transmissions and read from memory
✓ Some very playful challenges (Satellite, Large Cells, Many users)