Enabling "Aggressive" Non-Orthogonal Transmissions for Improved Connectivity in the Low SNR Regime

C. Jayawardena, G. N. Katsaros and K. Nikitopoulos
University of Surrey
Guildford, UK

ABSTRACT

Emerging applications such as direct-to-device satellite access, large-scale sensing, and connectivity in contested or infrastructure-denied environments must serve large populations of power-limited devices whose signals arrive at very low signal-to-noise ratios (SNRs), often with few or no degrees of freedom available to separate users. Conventional orthogonal access, multi-user MIMO, and non-orthogonal multiple access (NOMA) all rely on separability in time, space, code, or received power, capping the number of devices that can share a time-frequency resource. Such challenges call for a fundamental rethinking of both the transmission strategy, to maximize capacity gains, and the receiver processing, to translate these gains into practical throughput and connectivity benefits. To this end, we show that aggressive non-orthogonal transmission, in which many users transmit simultaneously on the same frequency resources, well beyond the typical two-user NOMA limit, unlocks substantial, previously unexploited capacity gains at low SNRs, even with a single-antenna receiver. However, to the best of our knowledge, no practical receiver has been shown to operate under such aggressive overloading. In this work, we revisit optimal receiver-design principles and propose computationally efficient processing schemes that translate these theoretical gains into practical throughput. Over-the-air experiments on a software-defined radio testbed show that four users transmitting concurrently to a single-antenna receiver achieve up to 2× the throughput of orthogonal-access and conventional NOMA baselines. We further show that these gains enable support for up to 10 devices at a fixed quality-of-service target, a capability that grows with the number of receiver antennas.

CCS CONCEPTS

• **Networks** → **Wireless access networks**; **Mobile networks**; • **Hardware** → **Signal processing systems**.

KEYWORDS

Non-orthogonal multiple access, low-SNR communications, massive connectivity

ACM Reference Format:
C. Jayawardena, G. N. Katsaros and K. Nikitopoulos. 2026. Enabling "Aggressive" Non-Orthogonal Transmissions for Improved Connectivity in the Low SNR Regime. In *The 25th ACM Workshop on Hot Topics in Networks (HotNets '26), November 16-17, 2026, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3847290.3859970

1 INTRODUCTION

Future wireless systems will increasingly need to support large populations of transmitters whose individual signals arrive at the receiver at very low signal-to-noise ratios (SNRs). This operating regime arises when transmit power is highly constrained, propagation distances are exceptionally large, or the communication infrastructure cannot be positioned sufficiently close to the users. Representative examples include public safety, disaster response, and contested or infrastructure-denied operation, where many users must communicate over long distances without access to nearby base stations or relays [2, 12, 14]; large-scale sensing networks (e.g., Low Power- Wide Area Networks [16]), where dense populations of energy-constrained devices operate under stringent battery and transmit-power budgets [7, 14]; and direct-to-device satellite communications [14], where large numbers of power-limited handsets may share a wide-area beam while communicating over distances of hundreds or thousands of kilometers, often at single-digit received SNRs [7]. In these settings, low SNR operation is not merely the result of suboptimal network deployment that can be overcome through densification. Rather, it is an intrinsic consequence of the link geometry, device energy budget, and practical deployment constraints.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HotNets '26, November 16-17, 2026, Salt Lake City, UT, USA
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-3029-0/2026/11
https://doi.org/10.1145/3847290.3859970

The resulting challenge is therefore not simply to establish a reliable low-SNR link. It is to support a large number of such links simultaneously while using the available spectrum and receiver infrastructure efficiently. This raises a fundamental question: when every transmitter is forced to operate in the low-SNR regime, how should uplink transmissions and receiver processing be designed to maximize both throughput and the number of devices that can be supported?

The baseline approach to multi-user access has been orthogonalization (e.g., Time/Frequency Division Multiple Access), which isolates users to avoid interference entirely but strictly caps the number of devices a resource element can serve. As the demand to support more simultaneous transmitters grew, the industry evolved toward multi-user(MU) MIMO [11, 15, 23] and Non-Orthogonal Multiple Access (NOMA) [27]. Crucially, however, these techniques were fundamentally designed for the high-SNR, interference-limited regime. In that regime, capacity scales with the spatial degrees of freedom. Consequently, MU-MIMO relies on large antenna arrays to maintain spatial separability between streams [20]. Similarly, conventional NOMA superimposes users carefully managing the resulting interference, that as we discuss later is relying on large power disparities for successive interference cancellation [19, 28] or sparse codes that strictly bound the number of users per resource [4, 17, 26]. Because these architectures were designed to mitigate interference at high SNR, they deliberately avoid heavily superimposing user streams beyond the available spatial or code dimensions.

However, in the low-SNR regime, the system becomes noise-limited rather than interference-limited. As we show later in detail, classical MIMO capacity theory [24] demonstrates that when individual transmitters are heavily power-limited, spatial degrees of freedom are no longer necessarily the bottleneck. Instead, at low SNR, sum capacity grows nearly linearly with total received energy. Adding the received energy contributions of many individually weak transmitters can therefore increase aggregate throughput even when the number of active users substantially exceeds the conventional spatial, temporal, or code-domain dimensions of the system. This observation motivates designing systems that, instead of reserving separate resources for weak users or limiting superposition to preserve separability, permit substantially more devices to transmit within the same resource element at low SNR. We refer to this design principle as *aggressive non-orthogonality*, that is the deliberate superimposing user signals well beyond the limits of conventional NOMA in order to capture the capacity gains implied by classical theory in the low-SNR regime.

In this work, first we show that the low-SNR regime itself harbors a distinct and previously unexplored capacity growth opportunity, one that only aggressive non-orthogonal transmissions are uniquely positioned to exploit. Specifically, we show that serving more than two user equipment (UE)s per receive antenna, unlocks capacity gains that are inaccessible to both orthogonal multiple access schemes and conventional NOMA. Realizing these gains requires receiver architectures capable of handling such dense superpositions and to the best of our knowledge, no existing design meets this need. We therefore introduce novel processing architectures with distinct performance-complexity trade-offs, grounded in optimal receiver principles, that make these theoretical gains practically attainable. Moreover, also for the first time this work provides the over-the-air demonstration of the aggressive non orthogonal transmission gains using conventional software-defined radios and we show that four UEs transmitting to a single-antenna receiver can achieve up to twice the sum-spectral-efficiency obtained when the same UEs are served using time-division MA (TDMA). We further show how these spectral-efficiency improvements translate into the ability to support significantly more devices at a fixed quality of service (QoS), supporting up to 10× more devices than orthogonal access across in indicative QoS tiers (i.e., broadband data, voice, and text). Finally, we discuss the key challenges and future directions to enable aggressively non-orthogonal access systems to be deployed at scale.

2 THE UNEXPLORED CAPACITY GAINS

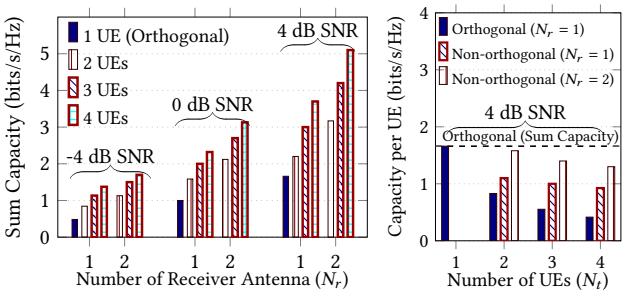
In this Section, we show that aggressive non-orthogonal transmissions offer substantial capacity gains in the low-SNR regime, and that these gains, though implied by classical capacity results, remain heavily unexploited by current systems. We consider uplink in which N_t single-antenna UEs transmit concurrently to a receiver equipped with N_r antennas. The $N_r \times 1$ received signal vector is given by

$$\mathbf{y} = \mathbf{H}\mathbf{s} + \mathbf{n}, \tag{1}$$

where $\mathbf{H}$ is the $N_r \times N_t$ channel matrix and $\mathbf{n}$ is the $N_r \times 1$ additive white Gaussian noise vector with per-element variance σ^2. The vector $\mathbf{s} \in \mathcal{S}^{N_t}$ contains the transmitted symbols, with $\mathbb{E}\{|s_i|^2\} = P_t$, where $\mathbb{E}\{\cdot\}$ denotes expectation. Each entry $H_{m,k}$ of $\mathbf{H}$ denotes the channel coefficient between the kth UE and the mth receive antenna. The capacity of (1) is

$$C = \log_2 \det\left(\mathbf{I}_{N_r} + \tfrac{P_t}{\sigma^2}\mathbf{H}\mathbf{H}^{\mathrm{H}}\right). \tag{2}$$

Although (2) is typically analyzed in the $N_r \geq N_t$ regime, its underlying mutual-information derivation makes no such requirement [24]. Yet the heavily underdetermined case ($N_t > N_r$) is often dismissed as infeasible because the resulting $\mathbf{H}$ is rank-deficient. Still, it is precisely in this aggressively non-orthogonal regime, and under low-SNR conditions, that the theoretical sum-capacity grows the most with the number of concurrent users, a gain that, as we demonstrate in this work, can be effectively realized in practice.



(a) Sum Capacity **(b) Capacity per UE**

Figure 1: Sum capacity versus the number of concurrent UEs for single- and two-antenna receivers.

To understand where the gains come from, note that each UE is power-limited, thus the only way to raise the total received power is to have more of them transmit at once. This is clearest for a single-antenna receiver ($N_r = 1$), where (2) reduces to

$$C = \log_2\left(1 + \frac{P_t}{\sigma^2}\sum_{k=1}^{N_t}|h_k|^2\right), \tag{3}$$

where the total received power is $P_t \sum_{k=1}^{N_t}|h_k|^2$, which scales directly with the sum of the individual channel gains. A common misconception here is that exploiting this gain requires the UEs to transmit the *same* signal coherently, so that their contributions add up in amplitude. They do not. Equation (3) is the *sum rate* of N_t UEs sending *independent* data, and it depends only on the aggregate received power. Each UE contributes its own channel gain to the sum irrespective of what the others send, so no alignment or coordination among UEs is needed. Capacity therefore grows monotonically with the number of concurrent UEs, most steeply at low SNR, where the logarithm is near-linear. Fig. 1 illustrates this capacity growth for a 3GPP TDL-A channel model with medium spatial correlation [1]. In particular, Fig. 1a shows the sum capacity, while Fig. 1b shows the capacity per UE. Notably, a single-antenna receiver cannot spatially separate the users, so in this case the gain stems from pooled received power alone. When $N_r > 1$, a spatial multiplexing gain adds to this power gain, further increasing capacity. As shown in Fig. 1b, the average capacity per UE decreases once the number of UEs exceeds N_r; nevertheless, both $N_r = 1$ and $N_r = 2$ retain significant gains over orthogonal multiple access (OMA). We assume TDMA, where each UE is active only $1/N_t$ of the total transmission slots, so the total transmitted power at any instant is P_t. In contrast, in non-orthogonal transmission, all UEs transmit simultaneously, each with power P_t. In Sec. 5, we quantify how these gains translate into an increased number of connected devices, depending on the quality-of-service target to be satisfied.

2.1 The Limitations of Existing Approaches

Realizing this gain requires *joint* processing of the superimposed signals. Existing practical non-orthogonal schemes generally avoid dense, unconstrained joint detection by imposing structure that simplifies receiver processing. Power-domain NOMA employs successive interference cancellation (SIC) [28]: users are decoded sequentially, with each stage treating all not-yet-decoded users as noise. The user decoded first therefore requires a received-power margin over the rest to attain a useful rate [19], and, at comparable target rates, effective SIC hinges on sufficient received-power disparity among the users. When the UEs are roughly power-balanced (e.g., in direct-to-device satellite scenarios), every stage faces comparable interference, and no decoding order yields a reliably decodable first user. Low SNR compounds the problem since high per-stage error rates mean that any misdecoded user corrupts the residual and propagates errors through all subsequent stages. Code-domain NOMA instead approximates the joint posterior by message passing on a factor graph [17, 26], which stays feasible only while the graph is sparse. This requires purpose-built sparse signatures that significantly cap the number of UEs per resource, and thus the very capacity gains we seek. In both cases the transmitted signal is shaped, by power disparity or by sparse signatures, to fit a suboptimal receiver, so the achievable gains are limited by the receiver rather than by the channel.

A separate line of work resolves dense collisions at a single-antenna receiver by exploiting incidental hardware offsets between low-cost transmitters as per-user separation signatures [6]. The offsets must be sufficiently large that each can be observed separately while preserving orthogonality. This shares the limitation of power- and code-domain NOMA that essentially depends on a resolvable transmit-side or hardware feature, so the achievable gain is bounded by the diversity of that feature rather than by the channel. In the following sections, we take a different approach. By revisiting optimal receiver-design principles, we show that, at low SNR, jointly processing heavily superimposed signals is not only feasible but often modest in computational cost, without requiring specifically designed transmit signals.

3 EXPLOITING THE CAPACITY GAINS

In this section, we revisit optimal receiver-design principles for aggressively non-orthogonal transmissions. We first establish the receiver-theoretic requirements for approaching the capacity gains of Section 2, independently of any particular detection algorithm. We then describe practical computational architectures for realizing these receivers, ranging from direct enumeration in small low-SNR systems to structured tree-search methods for larger configurations.

3.1 Maximum Likelihood (ML) Detection

Under additive white Gaussian noise and equiprobable transmitted symbols, maximum-likelihood (ML) detection is optimal in the sense of minimizing the probability of detection error. For the system model in (1), hard-decision ML detection is given by

$$\hat{\mathbf{s}}_{\mathrm{ML}} = \arg\min_{\mathbf{s}\in\mathcal{S}^{N_t}} \|\mathbf{y}-\mathbf{H}\mathbf{s}\|^2. \tag{4}$$

Crucially, (4) remains well defined even under aggressive non-orthogonal transmissions $N_t > N_r$, as long as distinct candidate vectors yield distinct metrics

$$d(\mathbf{s}) = \frac{1}{\sigma^2}\|\mathbf{y}-\mathbf{H}\mathbf{s}\|^2, \tag{5}$$

so that the minimizer is unique.

To see when this holds, consider the worst case for distinctness, where a single-antenna receiver ($N_r = 1$) with $N_t = 2$ users. ML detection reduces to $\hat{\mathbf{s}}_{\mathrm{ML}} = \arg\min_{\mathbf{s}\in\mathcal{S}^2} \left| y - h_1 s_1 - h_2 s_2 \right|^2$, and two candidates $\mathbf{s} \neq \mathbf{s}'$ are indistinguishable when $d(\mathbf{s}) = d(\mathbf{s}')$, which requires $h_1 s_1 + h_2 s_2 = h_1 s_1' + h_2 s_2'$. For example, with $h_1 = h_2$, the BPSK vectors $\mathbf{s} = [\,1, -1\,]^{\mathsf{T}}$ and $\mathbf{s}' = [\,-1, 1\,]^{\mathsf{T}}$ produce identical metrics for any noise realization and are therefore indistinguishable. While, exact collisions have extremely low probability under any continuous channel distribution, performance is governed by the minimum distance between noiseless received vectors. This distance takes small values with positive probability, and more often when users are spatially correlated. The resulting loss, however, is absorbed by the redundancy of the channel code, through soft-decision decoding or through iterative detection and decoding, both described in the following subsections.

3.2 Soft Information Computation for Receiver Processing

While hard-decision ML detection is optimal for uncoded systems, wireless standards such as 3GPP and Wi-Fi employ soft-input channel decoders, which require the detector to deliver soft information in the form of per-bit log-likelihood ratios (LLRs) to approach the minimum achievable bit-error probability. This is particularly important in the low SNR regime where per-symbol reliability is low and advanced forward error correction (FEC) is mandatory. In this context, the LLR of the ith bit of user l, denoted $b_{i,l}$, is defined as [22]

$$L_D(b_{i,l}) \triangleq \ln\!\left(\frac{P[\,b_{i,l}=1 \mid \mathbf{y},\mathbf{H}\,]}{P[\,b_{i,l}=0 \mid \mathbf{y},\mathbf{H}\,]}\right). \tag{6}$$

Applying the well-known max-log approximation, the LLR can instead be computed as [22]

$$L_D(b_{i,l}) \approx \mathrm{sign}\big(2b_{i,l}^{\mathrm{ML}}-1\big)\left(\min_{\mathbf{s}\in\mathcal{S}^{\bar{b}_{i,l}^{\mathrm{ML}}}} d(\mathbf{s}) - d(\hat{\mathbf{s}}_{\mathrm{ML}})\right), \tag{7}$$

where $b_{i,l}^{\mathrm{ML}}$ is the ith bit of user l in the bit label of the ML solution $\hat{\mathbf{s}}_{\mathrm{ML}}$, and $\mathcal{S}^{\bar{b}_{i,l}^{\mathrm{ML}}}$ denotes the subset of candidate vectors whose (i,l)th bit is the complement of $b_{i,l}^{\mathrm{ML}}$; the minimizer of this subset is referred to as the counter-hypothesis. Hence, computing (7) requires, in addition to the ML solution, one counter-hypothesis per transmitted bit.

3.3 Maximum-a-posteriori (MAP) Detection

The soft-output detector of Sec. 3.2 computes its LLRs in a single pass, treating all transmitted bits as equally likely a priori. However, the bits are constrained by the channel code, and only the decoder exploits these constraints. Feeding the decoder's soft information back to the detector as prior knowledge therefore sharpens the LLRs beyond what any single pass can achieve; a detector that incorporates such priors performs maximum a posteriori (MAP) detection rather than ML detection. Iterating this exchange approaches the performance of joint detection and decoding, which minimizes the bit-error probability and is optimal for coded MIMO transmission [8, 21]. The gain is most pronounced at low SNR, where the noisy observation alone yields the least information and the code constraints contribute the most.
Specifically, under the max-log approximation, the MAP detector computes the LLRs in (7) using the augmented metric

$$d(\mathbf{s}) = \frac{1}{\sigma^2}\,\|\mathbf{y}-\mathbf{H}\mathbf{s}\|^2 + I_{\mathrm{prior}}(\mathbf{s}), \tag{8}$$

where the prior term, which accounts for the a priori probability of the symbol vector $\mathbf{s}$, is given by

$$I_{\mathrm{prior}}(\mathbf{s}) = \frac{1}{2}\sum_{l=1}^{N_t}\sum_{i=1}^{\log_2|\mathcal{S}|}\Big(\big|L_A(b_{i,l})\big| - (2b_{i,l}-1)\,L_A(b_{i,l})\Big), \tag{9}$$

and $L_A(b_{i,l})$ denotes the a priori LLR of bit $b_{i,l}$ fed back by the channel decoder (e.g., an LDPC decoder) in the previous iteration. The LLRs are then computed as in (7), with the ML solution replaced by the MAP solution $\hat{\mathbf{s}}_{\mathrm{MAP}} = \arg\min_{\mathbf{s}\in\mathcal{S}^{N_t}} d(\mathbf{s})$ and the metric augmented as in (8).

To avoid feeding back already-known information, the detector passes only the extrinsic LLRs, $L_E(b_{i,l}) = L_D(b_{i,l}) - L_A(b_{i,l})$, to the decoder. Detector and decoder then exchange soft information over a small number of outer iterations, each pass refining the priors of the next.

3.4 Computational Realization: From Enumeration to Tree Search

Having specified the candidate-vector minimizations required by soft-output ML and MAP detection, we now consider how they can be computed efficiently.

3.4.1 ***Direct Enumeration in the Low-SNR Regime.*** For the low-order constellations and moderate numbers of user streams targeted in this work, direct enumeration can remain inexpensive. In particular, when $|\mathcal{S}| \le 4$ and $N_t \le 4$, all candidate vectors can be evaluated with modest computational cost. A 4×1 BPSK system, for example, contains only $2^4 = 16$ candidate vectors. Evaluating their metrics provides the global ML or MAP solution and all bit counter-hypotheses required by (7). Furthermore, the noiseless superpositions $\mathbf{Hs}$ can be precomputed once per channel-estimate update and reused for every received vector within the channel coherence interval. Only the distance calculations, together with the inexpensive prior penalties in the MAP case, must then be updated for each received signal. As N_t or $|\mathcal{S}|$ increases, however, the candidate count $|\mathcal{S}|^{N_t}$ grows exponentially. Structured search methods therefore become necessary for larger systems.

3.4.2 ***Regularized Tree Formulation.*** The exhaustive evaluation of Sec. 3.4.1 in the low-SNR regime is cheap. As N_t and $|\mathcal{S}|$ grow, however, the candidate set $\mathcal{S}^{N_t}$ expands exponentially and enumeration becomes impractical. Structured search recovers tractability by exploiting the problem geometry. Specifically, by applying the QR decomposition to the channel matrix, such problems can be reformulated as a tree search [21, 22], which reduces the complexity without sacrificing optimality.

In the overloaded regime $N_r < N_t$, the channel matrix $\mathbf{H}$ is rank-deficient and a conventional QR decomposition does not exist in order to support tree-search algorithms that solve (7) exactly [22]. Following [25], we instead decompose the Tikhonov-regularized matrix

$$\bar{\mathbf{H}} \triangleq \begin{bmatrix} \mathbf{H} \\ \lambda \mathbf{I}_{N_t} \end{bmatrix} = \bar{\mathbf{Q}}\mathbf{R} = \begin{bmatrix} \mathbf{Q}_1 \\ \mathbf{Q}_2 \end{bmatrix} \mathbf{R}, \tag{10}$$

with regularization parameter $\lambda = \sigma/\sqrt{P_t}$, where $\mathbf{Q}_1$ is $N_r \times N_t$ and $\mathbf{R}$ is $N_t \times N_t$ upper triangular. The metric $d(\mathbf{s})$ can then be expressed as

$$d(\mathbf{s}) = \frac{1}{\sigma^2}\|\tilde{\mathbf{y}} - \mathbf{R}\mathbf{s}\|^2 - \frac{\lambda^2}{\sigma^2}\|\mathbf{s}\|^2, \tag{11}$$

where $\tilde{\mathbf{y}} = \mathbf{Q}_1^H \mathbf{y}$ is an effective $N_t \times 1$ observation vector.

Owing to the upper-triangular structure of $\mathbf{R}$, the minimization problems in (7) can now map onto a tree search of height N_t and branching factor $|\mathcal{S}|$, and $d(\mathbf{s})$ can be accumulated recursively, one level at a time. Let $\mathbf{s}_k = [\, s_k, \ldots, s_{N_t} \,]^{\mathsf{T}}$ denote the partial symbol vector with partial metric $d(\mathbf{s}_k)$. Then, for $k = N_t, \ldots, 1$,

$$d(\mathbf{s}_k) = d(\mathbf{s}_{k+1}) + \frac{1}{\sigma^2}\Big| \tilde{y}_k - \sum_{l=k}^{N_t} R_{k,l}\, s_l \Big|^2 - \frac{\lambda^2}{\sigma^2}\,|s_k|^2, \tag{12}$$

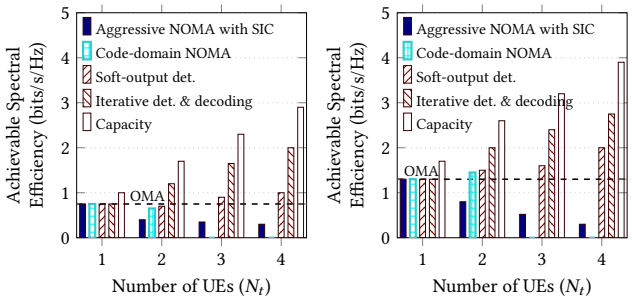
initialized with $d(\mathbf{s}_{N_t+1}) = 0$, so that $d(\mathbf{s}) = d(\mathbf{s}_1)$.

3.4.3 ***Incorporating MAP Priors.*** The MAP detection problem can be cast as the same tree search as in Sec. 3.4.2, applying the regularized decomposition (10); the only modification is that the recursive metric now incorporates the *a priori* information provided by the channel decoder. Specifically, for $k = N_t, \ldots, 1$,

$$d(\mathbf{s}_k) = d(\mathbf{s}_{k+1}) + \frac{1}{\sigma^2}\Big| \tilde{y}_k - \sum_{l=k}^{N_t} R_{k,l}\, s_l \Big|^2 + I_{\text{prior}}(s_k) - \frac{\lambda^2}{\sigma^2}\,|s_k|^2, \tag{13}$$

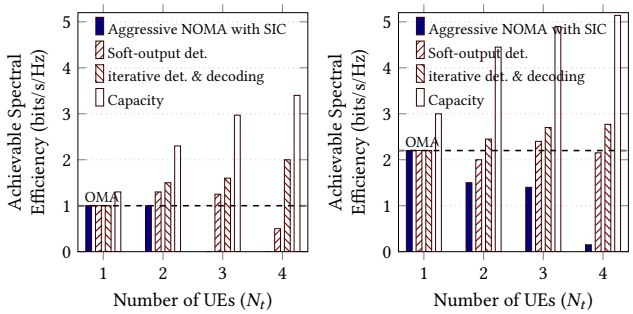
The prior adds only a per-symbol penalty to the recursion, so each MAP iteration costs essentially the same as one soft-output pass of Sec. 3.4.2; the only overhead is repeating that pass across a few outer iterations. In the low-order constellations and moderate user counts of the low-SNR regime, this per-iteration cost is modest, so full MAP detection remains practical. Realizing it efficiently for larger systems and denser constellations remains an open challenge.

3.4.4 ***Search Algorithms and Scalability.*** The regularized formulation above defines the tree and its recursive path metric, but does not prescribe how the tree should be explored. In principle, evaluating every leaf recovers the same ML or MAP minima as direct enumeration. Tree-search algorithms instead reduce the number of candidate vectors that must be visited by pruning unlikely partial paths or retaining only a restricted set of promising candidates. For example, the single-tree-search sphere decoder of [22] uses the recursive metric in (11) or (13) to compute the constrained minima required for the max-log LLRs in (7). It can recover the exact soft-output solution while visiting only part of the tree. Its traversal, however, is inherently sequential, and its complexity and latency depend strongly on the channel realization and noise conditions. The soft fixed-complexity sphere decoder [3, 5] instead expands a predetermined set of paths, enabling parallel execution and deterministic complexity. Closely approximating the LLRs may nevertheless require a large worst-case path list. Thus, tree search avoids the exhaustive evaluation of all $|\mathcal{S}|^{N_t}$ candidate vectors, but its computational cost still grows rapidly with the number of streams and modulation order. This motivates search strategies that identify a small set of high-quality candidate paths while retaining sufficient counter-hypotheses for accurate soft-output detection, as discussed in Sec. 5.



(a) 0 dB SNR (b) 4 dB SNR

Figure 2: Trace-driven evaluation of spectral efficiency as a function of the number of UEs ($N_r = 1$).



(a) 4 dB SNR (b) 10 dB SNR

Figure 3: Over-the-air evaluation of spectral efficiency as a function of the number of UEs ($N_r = 1$).

Adaptation to system dynamics: The receiver configuration depends only on the current channel estimates and the scheduling decision, i.e., the set of active UEs and their modulation orders. It is therefore updated only when a new channel estimate is obtained, once per coherence interval, or when the scheduler changes the set of active UEs. Variations in traffic demand and UE resources are handled by conventional scheduling and appropriate link adaptation. Each update recomputes only the channel-dependent preprocessing, i.e., either the precomputed signal superpositions or the QR decomposition, with complexities of $O(|\mathcal{S}|^{N_t})$ and $O(N_t^3)$, respectively. This cost is incurred once and reused for all symbols in the coherence interval, so the per-symbol overhead is small. The dominant factor is thus the detection complexity as a function of the number of concurrently scheduled UEs, analyzed in Section 5.

4 OVER-THE-AIR EVALUATION OF GAINS

In this Section, we validate that the gains discussed in Sec. 2 can practically translate into tangible throughput and connectivity gain in over the air environments. The testbed comprises of five NI USRP X310 (four single-antenna transmitters and one single-antenna receiver). Transmissions use a 3.4 GHz carrier at 40 MS/s, with a 128-point FFT, a 32-sample cyclic prefix, and six pilot subcarriers[1]. Reported SNR is the per-user average estimated from the pilots. Measurements span four locations and two modes. First, captured channels are replayed offline in MATLAB, so that all schemes, modulations, and user configurations are compared under identical channel realizations. Second, end-to-end over-the-air (OTA) decoding verifies that the gains persist under hardware non-idealities.

4.1 Results & Discussions

Figure 2 reports trace-driven achievable sum spectral efficiency against the number of concurrent UEs, at 0 and 4 dB SNR, alongside the sum capacity and the OMA (TDMA) reference, for a single active receive antenna. First, as UEs are added, the sum spectral efficiency climbs steadily, tracking the capacity curve, while exceeding the OMA reference even by over 2 times. The two conventional non-orthogonal baselines fail to capture this. Aggressive NOMA with SIC collapses once some users have comparable received powers, since SIC relies on large power disparities to separate them, while code-domain NOMA supports only a few concurrent users before its signature space is exhausted. Soft-output detection (Sec. 3.2) alone already yields significant gains, and adding iterations between detector and decoder, as described in Sec. 3.3, pushes achievable efficiency further still, reaching 70% of capacity with four streams on a single receive antenna and 2× the TDMA throughput. The end-to-end OTA experiments shown in Fig. 3 validate these gains and follow a similar trend (i.e., roughly 2× over OMA). However, the gap to capacity widens owing to additional quantization noise as well as synchronization and estimation errors. Moreover, as discussed in Sec. 2 and confirmed by Fig. 5b, the gain of adding further UEs diminishes at higher SNRs. Moreover, in our evaluation all concurrent UEs transmit at a common rate, whereas the capacity of each link depends on its own channel realization, so per-UE rate adaptation would increase sum throughput beyond what we report. Iterative detection and decoding approximates joint detection and decoding [18], whose complexity is prohibitive even for small N_t. Both factors contribute to the remaining gap to capacity.

5 FUTURE DIRECTIONS

The spectral-efficiency gains validated in Sec. 4 translate into connectivity gains for a given target QoS. In particular, the capacity in Fig. 1b and the subsequent spectral-efficiency evaluation imply a fundamental trade-off in aggressive non-orthogonal scenarios ($N_t > N_r$): the aggregate rate is shared among the supported users, so the lower the per-UE QoS

[1]Payloads are LDPC-encoded at rates 1/2, 2/3, 3/4, or 5/6 and supported modulation orders of BPSK, QPSK, 16-QAM, or 64-QAM.

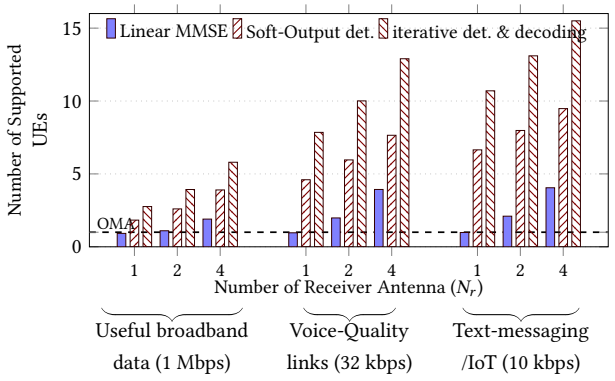


Figure 4: Number of UEs supported at a fixed QoS target for 1, 2, and 4 receive antennas, showing how antenna-driven spectral-efficiency gains translate directly into increased connectivity.

requirement, the larger the number of UEs that can be accommodated. Accordingly, Fig. 4 evaluates the number of supported UEs for three QoS categories, namely mobile broadband, voice, and text messaging[2], and compares the connectivity achieved by OMA, MIMO with MMSE detection, soft-output detection, and iterative detection and decoding. We show that iterative detection and decoding supports up to 10× as many concurrent UEs as orthogonal access at a fixed QoS, using only a single receive antenna.

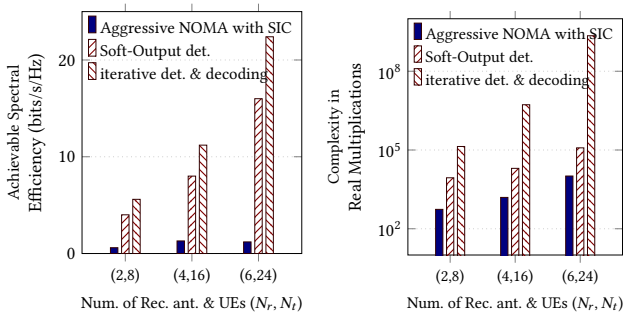
However, as shown in Table 1, the complexity of the advanced non-linear processing schemes that can address the receiver processing challenges discussed in Sec. 3 scales with the number of UEs (N_t). In this direction, recent non-linear processing frameworks [9, 13] identify only the most promising tree paths in a complexity-efficient and parallelizable manner, achieving real-time performance even in software, without dedicated hardware acceleration. Still, closely approaching the capacity limits would require joint detection and decoding architectures with a further improved performance–complexity tradeoff. Fig. [5] shows how the sum spectral efficiency and the complexity scale as the number of UEs increases. We consider 8, 16, and 24 UEs served by a receiver with 2, 4, and 6 antennas, respectively, so that the overloading factor remains fixed at four UEs per antenna. The sum spectral efficiency increases approximately linearly with the number of UEs when the ratio of UEs to receive antennas is fixed, whereas the complexity of iterative detection and decoding grows substantially. FlexCore [9] is used to solve the tree search for soft-output detection, whereas MAP detection employs [21]. As discussed in Sec. 3.4.4, the high complexity of iterative detection and decoding for large

[2]Broadband assumes an allocation of three resource blocks (1.08 MHz) per user at 30 kHz subcarrier spacing, whereas voice and text services assume the minimum allocation of one resource block (360 kHz); all services are evaluated at an SNR of 4 dB employing 3GPP TDL-A channel model [1].

Table 1: Receiver processing complexity scaling.

Scheme	Complexity
Capable of aggressive non-orthogonality	
MAP Detection-Decoding	$O(N_r N_t \lvert\mathcal{S}\rvert^{N_t} I_{out})$ (Sec. 3.3)
ML Detection	$O(N_r N_t \lvert\mathcal{S}\rvert^{N_t})$ (Sec. 3.1)
Sphere Decoder	$O(\lvert\mathcal{S}\rvert^{N_t})$, at low SNR [10]
FCSD	$O(N_t^3 + \lvert\mathcal{S}\rvert^{p} N_t^2 + N_t N_r)$ [3]
Flexcore	$O(N_t^3 + K N_t^2 + N_t N_r)$ [9]
Limited to conventional non-orthogonal access	
NOMA w/ SIC	$O(N_t^3 + N_t^2 + N_t N_r)$ [25, 28]
NOMA w/ BP	$O(I_{in}\, N_r \lvert\mathcal{S}\rvert^{d})$ [4]

$\lvert\mathcal{S}\rvert$: constellation size. N_t: number of streams. N_r: number of receive antennas. I_{out}: detection-decoding iterations. I_{in}: belief-propagation iterations. p: full-search dimensions in FCSD. K: candidates retained by Flexcore. d: streams superimposed per resource element, e.g., $d = 3$ when N_t(overloading factor) ≤ 2.



(a) Spectral Efficiency **(b) Complexity**

Figure 5: Sum spectral efficiency versus detection complexity for an increasing number of UEs.

numbers of UEs motivates future work on lower-complexity detection schemes.

All the receiver processing schemes discussed in this work rely on channel estimates. In standards-based systems, channel estimates for concurrent MU transmissions, such as in MU-MIMO, are obtained orthogonally; in time, in frequency, or through orthogonal sequences. Although this overhead is only around 5% for the four-user scenarios mainly considered in this work, the number of concurrently supported users remains limited by the available orthogonal pilot resources. For example, 3GPP 5G NR Release 18 supports a maximum of 24 orthogonal demodulation reference signal (DMRS) ports, double the 12 of Release 17. Since such scaling cannot continue indefinitely, resource-efficient channel estimation that supports larger numbers of UEs is a key direction for future releases and research.

Acknowledgments. This work was supported by the SNS JU under the EU's Horizon Europe programme, Grant Agreement No. 101290994 (SHARC).

REFERENCES

[1] 3GPP. 2020. *Study on Channel Model for Frequencies from 0.5 to 100 GHz.* Technical Report (TR) 38.901. 3rd Generation Partnership Project (3GPP). https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3173 Version 16.1.0.

[2] Gianmarco Baldini, Stan Karanasios, David Allen, and Fabrizio Vergari. 2013. Survey of Wireless Communication Technologies for Public Safety. *IEEE Communications Surveys & Tutorials* 16 (09 2013), 619–641. https://doi.org/10.1109/SURV.2013.082713.00034

[3] Luis G. Barbero, T. Ratnarajah, and Colin Cowan. 2008. A Low-Complexity Soft-MIMO Detector Based on the Fixed-Complexity Sphere Decoder. In *proc. IEEE ICASSP*. 2669–2672. https://doi.org/10.1109/ICASSP.2008.4518198

[4] Alireza Bayesteh, Hosein Nikopour, Mahmoud Taherzadeh, Hadi Baligh, and Jianglei Ma. 2015. Low Complexity Techniques for SCMA Detection. In *Proc. of IEEE Globecom Workshops (GC Wkshps)*. 1–6. https://doi.org/10.1109/GLOCOMW.2015.7414184

[5] Ya-Xin Dai, Shih-Jie Jhang, Yen-Ming Chen, Sheng-Ping Lan, and Yeong-Luh Ueng. 2022. An Efficient Soft Output MIMO Detector Architecture Considering High-order Modulations. In *Proc. of IEEE ACSSC*. 623–627. https://doi.org/10.1109/IEEECONF56349.2022.10051955

[6] Rashad Eletreby, Diana Zhang, Swarun Kumar, and Osman Yağan. 2017. Empowering Low-Power Wide Area Networks in Urban Settings. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication* (Los Angeles, CA, USA) *(SIGCOMM '17)*. Association for Computing Machinery, New York, NY, USA, 309–321. https://doi.org/10.1145/3098822.3098845

[7] Tim Farrar. 2025. *Satellite Direct-to-Device: A Supplement for Terrestrial Cell Coverage (White Paper)*. Technical Report. Telecom, Media and Finance Associates, Inc.

[8] B.M. Hochwald and S. ten Brink. 2003. Achieving near-capacity on a multiple-antenna channel. *IEEE Transactions on Communications* 51, 3 (2003), 389–399. https://doi.org/10.1109/TCOMM.2003.809789

[9] Christopher Husmann, Georgios Georgis, Konstantinos Nikitopoulos, and Kyle Jamieson. 2017. Flexcore: massively parallel and flexible processing for large MIMO access points *(NSDI'17)*. USENIX Association, USA, 197–211.

[10] J. Jalden and B. Ottersten. 2005. On the complexity of sphere decoding in digital communications. *IEEE Transactions on Signal Processing* 53, 4 (2005), 1474–1484. https://doi.org/10.1109/TSP.2005.843746

[11] Huangping Jin, Kunpeng Liu, Min Zhang, Leiming Zhang, Gilwon Lee, Emad N. Farag, Dalin Zhu, Eko Onggosanusi, Mansoor Shafi, and Harsh Tataria. 2023. Massive MIMO Evolution Toward 3GPP Release 18. *IEEE Journal on Selected Areas in Communications* 41, 6 (2023), 1635–1654. https://doi.org/10.1109/JSAC.2023.3273768

[12] Abhaykumar Kumbhar, Farshad Koohifar, İsmail Güvenç, and Bruce Mueller. 2017. A Survey on Legacy and Emerging Technologies for Public Safety Communications. *IEEE Communications Surveys & Tutorials* 19, 1 (2017), 97–124. https://doi.org/10.1109/COMST.2016.2612223

[13] Konstantinos Nikitopoulos, Marcin Filo, George N Katsaros, Chathura Jayawardena, and Rahim Tafazolli. 2023. MU-MIMO, Open-RAN PHY with Linear and Massively Parallelizable Non-Linear Processing. In *MobiCom*. 1–3.

[14] Hannaneh B. Pasandi, Juan A. Fraire, Sylvia Ratnasamy, and Herve Rivano. 2024. A Survey on Direct-to-Device Satellite Communications: Advances, Challenges, and Prospects. In *Proc. of the International Workshop on LEO Networking and Communication* (Washington, DC, USA) *(LEO-NET '24)*. Association for Computing Machinery, New York, NY, USA, 7–12. https://doi.org/10.1145/3697253.3697265

[15] K. Sai Praneeth, Jeeva Keshav Sattianarayanin, Sai Prasanth Kotturi, Rohit Singh, Aravind Gundlapalle, Siddharth Singh, and Radha Krishna Ganti. 2025. Design of a 32 Channel 5G NR MIMO Base Station. In *Proc. of the ACM Workshop on Open and AI RAN (OpenRan '25)*. Association for Computing Machinery, New York, NY, USA, 43–49. https://doi.org/10.1145/3737900.3770173

[16] Usman Raza, Parag Kulkarni, and Mahesh Sooriyabandara. 2017. Low Power Wide Area Networks: An Overview. *Commun. Surveys Tuts.* 19, 2 (April 2017), 855–873. https://doi.org/10.1109/COMST.2017.2652320

[17] Manel Rebhi, Kais Hassan, Kosai Raoof, and Pascal Chargé. 2021. Sparse Code Multiple Access: Potentials and Challenges. *IEEE Open Journal of the Communications Society* 2 (2021), 1205–1238. https://doi.org/10.1109/OJCOMS.2021.3081166

[18] Lars Schmitt and Heinrich Meyr. 2006. Turbo Decoding as an Approximative Iterative Solution to Maximum Likelihood Sequence Detection. In *Proc. of International Symposium on Turbo Codes & Related Topics; 6th International ITG-Conference on Source and Channel Coding*. 1–6.

[19] Souvik Sen, Naveen Santhapuri, Romit Roy Choudhury, and Srihari Nelakuditi. 2010. Successive interference cancellation: a back-of-the-envelope perspective. In *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks* (Monterey, California) *(Hotnets-IX)*. Association for Computing Machinery, New York, NY, USA, Article 17, 6 pages. https://doi.org/10.1145/1868447.1868464

[20] Clayton Shepard, Hang Yu, and Lin Zhong. 2013. ArgosV2: a flexible many-antenna research platform. In *Proc. of the Annual International Conference on Mobile Computing & Networking* (Miami, Florida, USA) *(MobiCom '13)*. Association for Computing Machinery, New York, NY, USA, 163–166. https://doi.org/10.1145/2500423.2505302

[21] C. Studer and H. Bolcskei. 2010. Soft Input Soft Output Single Tree-Search Sphere Decoding. *IEEE Trans. Inf. Theory* 56, 10 (Oct 2010), 4827–4842. https://doi.org/10.1109/TIT.2010.2059730

[22] C. Studer, A. Burg, and H. Bolcskei. 2008. Soft-output sphere decoding: algorithms and VLSI implementation. *IEEE J. Sel. Areas Commun.* 26, 2 (February 2008), 290–300.

[23] Tomoya Tandai, Hiroki Mori, Kiyoshi Toshimitsu, and Takahiro Kobayashi. 2009. An efficient uplink multiuser MIMO protocol in IEEE 802.11 WLANs. In *Proc. of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*. 1153–1157. https://doi.org/10.1109/PIMRC.2009.5450289

[24] Emre Telatar. 1999. Capacity of multi-antenna Gaussian channels. *European transactions on telecommunications* 10, 6 (1999), 585–595.

[25] D. Wubben, R. Bohnke, V. Kuhn, and K. D. Kammeyer. 2003. MMSE extension of V-BLAST based on sorted QR decomposition. In *Proc. of IEEE VTC*, Vol. 1. 508–512 Vol.1.

[26] Lisu Yu, Zilong Liu, Miaowen Wen, Donghong Cai, Shuping Dang, Yuhao Wang, and Pei Xiao. 2021. Sparse Code Multiple Access for 6G Wireless Communication Networks: Recent Advances and Future Directions. *IEEE Commun. Stand. Mag.* 5, 2 (2021), 92–99. https://doi.org/10.1109/MCOMSTD.001.2000049

[27] Yifei Yuan, Zhifeng Yuan, and Li Tian. 2020. 5G Non-Orthogonal Multiple Access Study in 3GPP. *IEEE Communications Magazine* 58, 7 (2020), 90–96. https://doi.org/10.1109/MCOM.001.1900450

[28] Volkan Özduran, Mohammadali Mohammadi, Nikolaos Nomikos, Imran Shafique Ansari, and Panagiotis Trakadas. 2024. On the performance of uplink power-domain NOMA with imperfect CSI and SIC in 6G networks. *Journal of Communications and Networks* 26, 4 (2024), 445–460. https://doi.org/10.23919/JCN.2024.000039