Scaling a Reliable Distributed System

Rachid Guerraoui* EPFL

Alexandre Maurer† EPFL

Abstract

We consider the problem of *reliably* connecting an arbitrarily large set of computers (nodes) with communication channels. Reliability means here the ability, for any two nodes, to remain connected (i.e., their ability to communicate) with probability at least μ, despite the very fact that every other node or channel has an independent probability λ of failing. A simple solution to the problem consists in connecting *every* pair of nodes with several channels. This solution however does not scale: the number of connections per node (degree) would not be bounded by a constant.

We address the following question: is it possible to reliably connect an arbitrarily large number n of nodes with a bounded degree? This problem is non-trivial, as the level of redundancy implied by reliability is apparently incompatible with a bounded degree. In this paper, we show that, may be surprisingly, the answer to this problem is positive. We show how to build a graph to reliably connect n nodes while preserving a bounded degree.

We first address a weak version of the problem, where we allow to add *intermediary* nodes (that are not necessarily reliably connected to the others), provided that their number is *linear* in the total number of nodes reliably connected. To solve the weaker problem, we define a fractal graph that ensures constant reliability at any distance, and combine it with a tree-like graph to reliably connect an arbitrary set of nodes. Then, to solve the strong version of the problem (without intermediary nodes), we split the n nodes to connect into several subsets, and reliably connect each pair of subsets with an instance of the previous graph containing at most n nodes. The final graph is obtained by merging all these instances together. The linearity property of the weak problem ensures that the number of graphs we merge is bounded by a constant, which guarantees a bounded degree. Interestingly, the resulting graph has an optimal diameter: it is logarithmic in n.

Whilst we focus on crash-stop failures for presentation simplicity, we also show how our solution can be generalized to tolerate *Byzantine* (malicious) failures, by increasing the level of redundancy and performing majority votes at several levels of the graph.

1 Introduction

With the fast development of communication networks, more and more computers are getting connected. The growth of the size of modern networks seems to be even exceeding Moore's Law [22]. This is in particular the case for data centers handling massive data storage for cloud computing [5, 4], as well as industrial and research simulations. We talk about 60,000 cores for the Human Brain Project [1] and over 100,000 for the CERN data center [2]. Companies like Google and Microsoft today have data centers of more than one million servers [3].

We consider the problem of *reliably* connecting a set of computers (we say *nodes*) with communication channels. The reliability criteria we consider is the following: assuming that each node or

*rachid.guerraoui@epfl.ch

†alexandre.maurer@epfl.ch

channel has an independent probability λ to crash[1], any two (non-crashed) nodes must be able to communicate with probability at least μ. We call this the *communication probability*[2].

A simple solution to the problem consists in building a "complete graph", that is, to add one or several communication channels between any two nodes. This solution is clearly not scalable: as each node needs to be connected to all others, the node degree (i.e., the number of channels connected to a given node) explodes. In practice, we can only connect a finite number of channels to a given node. In order to scale, the node degree must be *bounded* by a constant.

Many network topologies were proposed to reliably connect a large number of nodes with a "reasonable" degree [15, 6, 20, 11, 12, 19, 7, 9]. However, all proposed approaches were empirical and have only been experimented through simulations: their performances were evaluated for a specific number of nodes. In fact, if we consider the asymptotic behavior of their proposed graphs (i.e., when the number of nodes grows larger and larger), either the communication probability approaches zero, or the maximal degree approaches infinity.

We study for the first time this problem theoretically, and we address the following question: is it possible to connect an arbitrarily large number of nodes n, while achieving any desired level of reliability *and* preserving a bounded degree? We call this problem the RBD (**R**eliable **B**ounded **D**egree) problem. (We give the precise definition of the problem in the paper.)

We keep the setting voluntarily simple here: all nodes and channels have the same status and the same probability of failure (in practice, this network could represent the backbone of the actual network). We seek no optimization of the degree of the network and focus on the feasibility of the problem, first in the context of crash failures.

At first glance, the answer to the RBD problem seems to be negative. Indeed, consider a graph of n nodes with a bounded degree. When n increases, the diameter of the graph also increases: some pairs of nodes become more and more distant from each other, inevitably dragging down the communication probability. To compensate for this loss of reliability, the natural solution is to add redundant paths between any pair of (distant) nodes. However, the number of parallel paths is bounded by the maximal degree here, while the network diameter keeps increasing with n. Clearly, for a sufficiently large n, the loss of reliability cannot be compensated by a bounded number of parallel paths. The trade-off seems impossible to circumvent.

In this paper, we show that, may be surprisingly, the answer to the RBD problem turns out to be positive. More precisely, we provide a solution to this problem: for any number of nodes n, we show how to build a graph G_n that ensures arbitrarily high reliability while preserving a bounded degree.

We proceed in two main parts, each one containing several steps.

1. We first solve a *weak* version of the problem we call the **W**eak **RBD** (WRBD) problem, which we believe is interesting in its own right. The goal is to reliably connect m nodes with a graph W_m of bounded degree. The difference with the RBD problem is that it is allowed to add *intermediary* nodes between these m nodes (that are not necessarily reliably connected to the rest), provided that their number is at most linear in m – that is, at most Cm, where C is a constant. The key idea of our solution to this problem is to define a *fractal* graph that ensures a constant communication probability between any two given nodes (independently of their distance) with a bounded degree. This fractal definition enables us to express the communication probability as a *convergent sequence*. The fractal graph can be described as a

[1]Here, "crash" refers to the classical "crash-stop" model [21, 14] .

[2]This criteria should not be confused with the following: "the whole graph should remain connected with probability λ". Indeed, this second criteria is impossible to satisfy: as the node degree must be bounded by a constant, when the size of the network increases, the probability that all channels surrounding some node crash approaches 1. Therefore, it is impossible to have a lower bound on the probability that the whole graph remains connected. For this reason, we consider a less restrictive criteria, that is: the probability that p and q are connected (where p and q can be any nodes).

floor graph: a graph where the nodes are divided into floors, each floor being only connected to adjacent floors. Then, we define a tree-like floor graph connecting m nodes, and combine both graphs "floor by floor" to reliably connect the m nodes. The different speed of growth between the floors of these two graphs enables to preserve a linear number of intermediary nodes: overall, the number of intermediary nodes is divided by two every two floors, and their total number is therefore a convergent sum.

2. We then use the solution to the WRBD problem to solve our seemingly stronger RBD problem (i.e., reliably connecting n nodes *without* intermediary nodes). The idea is to combine several instances of a WRBD graph W_m, each instance reliably connecting a smaller number of nodes $m \leq n$, and to make their intermediary nodes "disappear" by merging them with other nodes. More precisely, let m be an integer sufficiently small so that W_m contains at most n nodes in total (including the intermediary nodes). We divide the n nodes into several subsets, and connect each pair of subsets with an instance of W_m. For each instance of W_m, we merge the intermediary nodes with the $n - m$ other nodes. According to the linearity property of the WRBD problem, the number of subsets into which n should be divided is bounded. Then, even after merging all the instances of W_m with the n nodes, the degree remains bounded. Interestingly, the resulting solution has an optimal (logarithmic) diameter.

In the paper, we show how to extend our result to Byzantine failures (when the failed components, i.e., nodes or channels, have an arbitrary malicious behavior). Basically, assuming a failure rate $\lambda < 0.5$ (which is necessary here), we can still solve the RBD problem, even with Byzantine failures, by increasing the level of redundacy and adding several layers of majority votes.

The paper is organized as follows. In Section 2, we define the WRBD problem (the weak version of the problem) as well as the RBD problem itself. In Section 3, we define a graph W_m that solves the WRBD problem, and prove its correctness. In Section 4, we define a graph G_n that solves the RBD problem, and prove its correctness. In Section 5, we show that the diameter of our solution is optimal. In Section 6, we discuss the extension to Byzantine failures. We conclude in Section 7 by discussing some related works and possible extensions.

2 The problems

In this section, we state some definitions, then define the WRBD and RBD problems.

Definitions. Let $\lambda \in]0,1[$ and $\mu \in]0,1[$ be any two arbitrary values: λ represents the (independent) probability of failure of each node or channel, and μ the desired communication probability between each pair of nodes. We define these notions below.

A graph is a tuple $G = (V, E)$ where V is the set of *nodes* and E is the set of *channels*. E is a set of pairs of nodes $\{p, q\} \subseteq V$. In this paper, E is a set with repetition: for two nodes p and q, it is possible to have multiple channels between p and q.

A *component* of a graph G is any node or channel of G. Each component of G can be either *alive* (functional) or *crashed* (failed). An *alive path* is a sequence of nodes $(p_1, \ldots, p_m)$ such that, $\forall i \in \{1, \ldots, m\}$, p_i is alive, and $\forall i \in \{1, \ldots, m-1\}$, there exists an alive channel $\{p_i, p_{i+1}\}$. Two nodes p and q are *connected* if there exists an alive path $(p_1, \ldots, p_m)$ such that $p_1 = p$ and $p_m = q$.

The *communication probability* of two nodes p and q is the probability that p and q are connected (according to the definition just above) when p and q are alive and any other component is crashed with an independent probability λ. We say that p and q are *reliably connected* if the communication probability between p and q is at least μ.

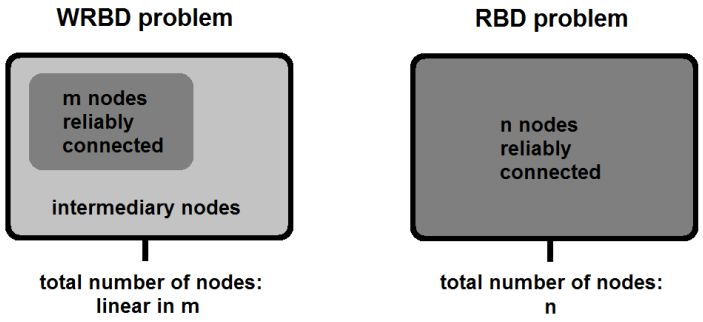

Figure 1: The difference between the WRBD and RBD problems.

WRBD. The WRBD (Weak Reliable Bounded Degree) problem consists in finding a graph W_m (with a parameter $m \geq 2$) satisfying the three following requirements:

1. **Reliability.** $\forall m \geq 2$, there exists a set S_m of nodes of W_m such that $|S_m| = m$ and, for any two nodes p and q of S_m, p and q are reliably connected (in other words, at least m nodes of W_m are reliably connected).
2. **Bounded degree.** There exists a constant Δ such that, $\forall m \geq 2$, the maximal degree of W_m is at most Δ (that is, each node of W_m is connected to at most Δ channels).
3. **Linear number of nodes.** There exists a constant C such that, $\forall m \geq 2$, the number of nodes of W_m is at most Cm (that is, the number of nodes is linear in m).

RBD. The RBD (Reliable Bounded Degree) problem consists in finding a graph G_n (with a parameter $n \geq 2$), *containing exactly n nodes* and satisfying the two following requirements:

1. **Reliability.** $\forall n \geq 2$, for any two nodes p and q of G_n, p and q are reliably connected.
2. **Bounded degree.** There exists a constant Δ such that, $\forall n \geq 2$, the maximal degree of G_n is at most Δ.

The difference between the WRBD and RBD problems is illustrated in Figure 1. In the WRBD problem, m nodes need to be reliably connected, and the total number of nodes is linear in m (i.e., at most Cm). In other words, in the WRBD problem, some nodes are not necessarily reliably connected to the others. We call them *intermediary nodes*. These intermediary nodes can represent routers which purpose is only to connect m computers reliably. In the RBD problem, there are no intermediary nodes: the n nodes of the graph need to be reliably connected.

3 Solving the WRBD problem

In this section, we define a graph W_m (3.1) and prove that it solves the WRBD problem (3.2).

3.1 A WRBD graph

Overview. For the motivation of the construction steps, please refer to the introduction. Here, we explain these steps in more details.

We first define the notion of *floor graph*. A floor graph is a graph where nodes are separated into several "floors", and where only nodes of two adjacent floors can be connected.

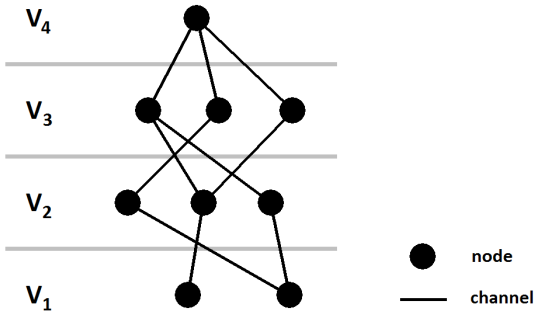

Figure 2: A floor graph of height $H = 4$.

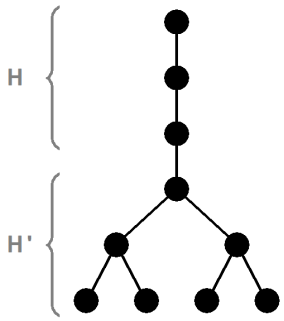

Figure 3: Structure of graph T_m.

Then, we define two floor graphs: T_m, which contains a binary tree connecting at least m nodes, and R_m, which is a "fractal" graph defined by induction. The fractal definition of R_m enables to preserve a constant communication probability between the first and last floor (independently of m) when $\lambda < 0.01$ (Lemma 1). We show how to overcome this "$\lambda < 0.01$" constraint below. Besides, R_m is defined so that the number of nodes doubles at most every 2 floors, which enables to preserve a linear number of nodes, as shown in Theorem 3. The number of floors of T_m is adjusted so that T_m and R_m have the same number of floors H_m.

Then, we define a graph X_m, which is a "floor by floor" product of T_m and R_m, and a graph Y_m, which puts two graphs X_m in parallel. Doing so ensures a constant communication probability between any two nodes of the first floor.

Finally, we make three transformations in order to reach any communication probability μ with any failure rate λ. First, we connect several graphs Y_m in parallel, in order to achieve any communication probability μ. Second, we replicate each node, in order to simulate a failure rate $\lambda < 0.01$ for each node. Third, we replicate each channel, in order to simulate a failure rate $\lambda < 0.01$ for each channel. The graph thus obtained is W_m.

Definitions. For any $m \geq 2$, let h_m be the smallest integer such that $2^{h_m-1} \geq m$. Let K_m be the smallest integer such that $2 + 4K_m \geq h_m$, and let $H_m = 2 + 4K_m$. Let α be the smallest integer such that $\alpha \geq 1$ and $0.5^\alpha \leq 1 - \mu$. Let β be the smallest integer such that $\beta \geq 1$ and $\lambda^\beta \leq 0.01$.

A *floor graph* of height H is a tuple $(V_1, \ldots, V_H, E)$ satisfying the three following conditions:

1. (V, E) is a graph with $V = \bigcup_{i \in \{1,\ldots,H\}} V_i$.
2. The sets V_i ("floors") are disjoint: $\forall \{i, j\} \subseteq \{1, \ldots, H\}$, $V_i \cap V_j = \emptyset$.
3. The channels only connect neighbor floors: $\forall \{p, q\} \in E$, if $p \in V_i$ and $q \in V_j$, then $|i - j| = 1$.

An example of a floor graph is given in Figure 2. By convention, in the following figures, V_1 always corresponds to the lower floor on the figure. We call V_1 the "first floor" and V_H the "last floor".

Graph T_m. We first define a tree-like floor graph of height H_m. Consider the floor graph represented in Figure 3: this graph is composed of a line of height $H = 3$ and of a binary tree of height $H' = 3$. In other words, $\forall i \in \{1, \ldots, H'\}$, the floor i contains 2^{i-1} nodes, and the H remaining floors contain each 1 node. Then, $\forall m \geq 2$, we define T_m as a similar graph with $H = H_m - h_m$ and $H' = h_m$.

Graph R_m. $\forall k \geq 0$, we first define a floor graph Q_k by induction. Let Q_0 be a floor graph of height 2 containing 2 nodes and 1 channel, as described in Figure 4. Then, $\forall k \geq 0$, Q_{k+1} is constructed with 2 instances of Q_k in parallel and 4 additional nodes, as described in Figure 4 (Q_{k+1} has 4 more floors than Q_k). We now define R_m as follows: $\forall m \geq 2$, $R_m = Q_{K_m}$.

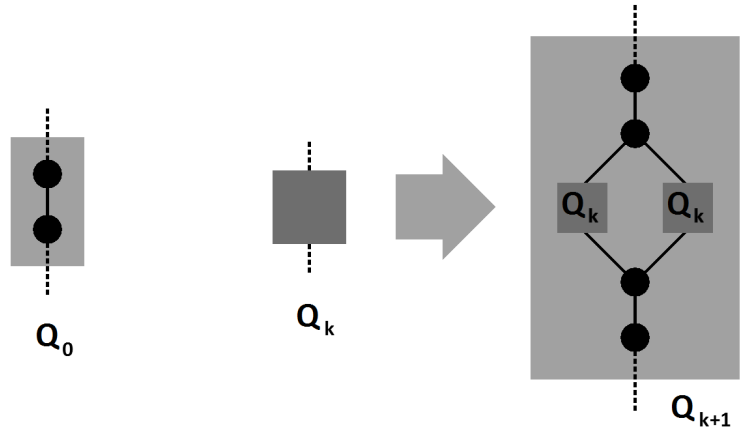

Figure 4: Construction (by induction) of fractal graph Q_k.

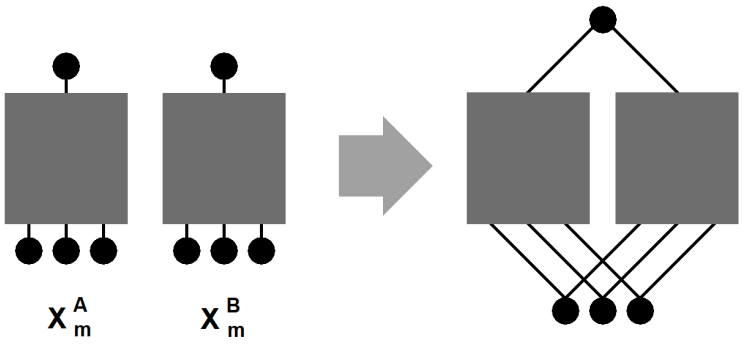

Figure 5: Construction of graph Y_m.

Graph X_m. $\forall m \geq 2$, T_m is a floor graph of height H_m, and R_m is a floor graph of height $2+4K_m = H_m$. As T_m and R_m are floor graphs, let $T_m = (V_1, \ldots, V_{H_m}, E)$ and $R_m = (V'_1, \ldots, V'_{H_m}, E')$. Then, $\forall m \geq 2$, we define the floor graph $X_m = (V^*_1, \ldots, V^*_{H_m}, E^*)$ as follows:

- $\forall i \in \{1, \ldots, H_m\}$, to each pair of nodes $(u, v) \in V_i \times V'_i$, we associate a unique node $p = f(u, v) \in V^*_i$ (thus, $|V^*_i| = |V_i||V'_i|$).
- Let $p = f(u, v)$ and $p' = f(u', v')$. Then, p and p' are neighbor in X_m if and only if u and u' (resp. v and v') are neighbors in T_m (resp. R_m).

Observe that, as the last floors of T_m and R_m contain 1 node, the last floor of X_m also contains 1 node.

Graph Y_m. $\forall m \geq 2$, we define the graph Y_m as follows: we consider two instances of X_m (X^A_m and X^B_m), we merge the nodes of their first floors, and we merge the nodes of their last floors. This is illustrated in Figure 5.

Graph W_m. $\forall m \geq 2$, the graph W_m is finally obtained by applying three successive transformations to Y_m:

1. **Transformation 1 (Network replication).** First, we connect α instances of Y_m by merging the nodes of their first floors. This is illustrated in Figure 6 for $\alpha = 3$.

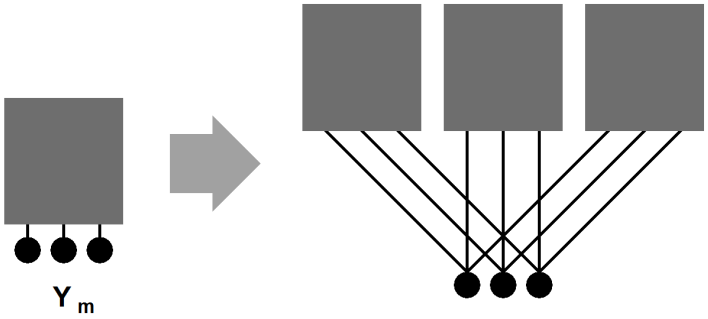

Figure 6: Transformation 1 (Network replication) with $\alpha = 3$.

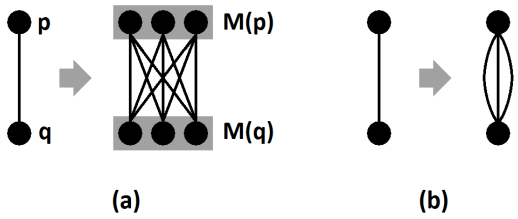

Figure 7: Transformations 2 (Node replication) and 3 (Channel replication) with $\beta = 3$.

2. **Transformation 2 (Node replication).** Second, we replace each node p by a set of β nodes $M(p)$. Then, for each channel $\{p, q\}$, we add a channel between each node of $M(p)$ and each node of $M(q)$ (see Figure 7-a).
3. **Transformation 3 (Channel replication).** Third, we replace each channel by β channels in parallel (see Figure 7-b).

3.2 Correctness

We prove that the graph W_m solves the WRBD problem. For this purpose, we prove the three properties of the WRBD problem: **Reliability**, **Bounded degree** and **Linear number of nodes**.

In Lemma 1, we show that, for a sufficiently small failure rate ($\lambda \leq 0.01$), we have a constant communication probability between the first and last floor of R_m (independently of m). To do so, we call P_k the probability that the first and last floor of Q_k are connected, then express P_{k+1} as a function of P_k (according to the inductive definition of Q_k). Then, we show that if $P_k \geq 0.8$, we also have $P_{k+1} \geq 0.8$. Thus, the communication probability between the first and last floor of Q_k (and thus, R_m) is at least 0.8.

In Lemma 2, we show that the first floor of W_m contains at least m nodes. Then, we consider that S_m is a subset of the first floor of W_m to prove the following property.

In Theorem 1, we prove the **Reliability** property. We first consider the case $\lambda \leq 0.01$ and $\mu \leq 0.5$ (in this case, $Y_m = W_m$). According to the definition of X_m and Y_m, any two nodes of S_m are connected to the last floor of Y_m by two graphs R_m. Thus, the result, according to Lemma 2. We then consider that λ and μ can have any value, and show that the 3 final transformations of 3.1 enable to simulate the previous situation where $\lambda \leq 0.01$ and $\mu \leq 0.5$.

In Theorem 2, we prove the **Bounded degree** property. As W_m is intentionally defined as a combination of graphs with a bounded degree, the property follows.

In Theorem 3, we prove the **Linear number of nodes** property. We use the fact that the number of nodes of T_m is divided by 2 every floor (starting from the first floor), while the number of nodes of

R_m at most doubles every 2 floors. Therefore, the number of nodes of X_m (which is the combination of T_m and R_m) is at least divided by 2 every 2 floors. Then, as $1 + 1/2 + 1/4 + 1/8 + \cdots \leq 2$, the number of nodes of X_m is linear in m, and so is the number of nodes of W_m.

Lemma 1. *If $\lambda \leq 0.01$, then $\forall m \geq 2$, the communication probability of the nodes of the first and last floor of R_m is at least 0.8.*

Proof. First, note that according to the definitions of Section 2, "p and q are connected with probability P" is a stronger property than "the communication probability of p and q is P", as the second one assumes that p and q are alive.

$\forall k \geq 0$, let p_k (resp. q_k) be the only node of the first (resp. last) floor of the graph Q_k. Let P_k be the probability that p_k and q_k are connected.

Let $k \geq 0$. Figure 4 shows how Q_{k+1} is constructed with 2 instances of Q_k and 10 additional components. Then, observe that p_{k+1} and q_{k+1} are connected in the following particular situation: the 10 additional components are all alive, and at least one of the two instances of Q_k has the nodes of its first and last floor connected (which happens with probability P_k). Therefore, $P_{k+1} \geq p(P_k)$, with $p(x) = (1-\lambda)^{10}(1-(1-x)^2)$.

The function $p(x)$ is increasing for $x \in [0.8, 1]$, $p(0.8) \in [0.8, 1]$ and $p(1) \in [0.8, 1]$. Therefore, $\forall x \in [0.8, 1]$, $p(x) \in [0.8, 1]$.

As Q_0 contains 3 components, $P_0 \geq (1-\lambda)^3$. Thus, as $\lambda \leq 0.01$, $P_0 \geq 0.8$ and $P_0 \in [0.8, 1]$. Therefore, by induction, $\forall k \geq 0$, $P_k \in [0.8, 1]$: p_k and q_k are connected with probability 0.8. Thus, as $R_m = Q_{K_m}$, the result follows. □

Lemma 2. *$\forall m \geq 2$, the first floor of W_m contains at least m nodes.*

Proof. Let $m \geq 2$. The first floor of T_m contains $2^{h_m - 1} \geq m$ nodes. Then, by definition of X_m, the first floor of X_m contains at least m nodes, and so does the first floor of Y_m. Thus, as the 3 final transformations of 3.1 can only increase the number of nodes of each floor, the first floor of W_m contains at least m nodes. □

Theorem 1. *$\forall m \geq 2$, there exists a set S_m of nodes of W_m such that $|S_m| = m$ and, for any two nodes p and q of S_m, p and q are reliably connected*

Proof. According to Lemma 2, $\forall m \geq 2$, let S_m be a set containing m nodes of the first floor of W_m.

Let $m \geq 2$, and let p and q be any two nodes of S_m. First, assume that $\lambda \leq 0.01$ and $\mu \leq 0.5$. Then, $\alpha = 1$ and $\beta = 1$, and according to the 3 final transformations of 3.1, W_m is identical to Y_m. Let a be a node of the first floor of X_m, and let b be the only node of the last floor of X_m. Let P_0 be the communication probability of a and b in X_m. Then, according to the definition of X_m, P_0 is at least the communication probability of the nodes of the first and last floor of R_m. Thus, according to Lemma 1, $P_0 \geq 0.8$.

As Y_m is formed by 2 instances of X_m, the communication probability of p and q is $P_1 \geq P_0^2(1-\lambda) \geq 0.5$ (as $P_0 \geq 0.8$ and $\lambda \leq 0.01$). Thus, as $\mu \leq 0.5$ here, $P_1 \geq \mu$, and p and q are reliably connected.

Now, we only assume that $\lambda \leq 0.01$ (μ can have any value in $]0, 1[$). Then, $\beta = 1$, and transformations 2 and 3 do not change anything. After transformation 1, the communication probability of p and q is $P_2 = 1 - (1 - P_1)^\alpha \geq 1 - 0.5^\alpha$ (as $P_1 \geq 0.5$). According to the definition of α, $0.5^\alpha \leq 1 - \mu$. Thus, $P_2 \geq \mu$, and p and q are reliably connected.

Finally, we consider that λ and μ can have any value in $]0, 1[$. Let us show that, after transformations 2 and 3, we reach a situation which is equivalent to the previous case where $\lambda \leq 0.01$.

Let Z_m be the graph after transformation 1. After transformation 2, each node u is replaced by a set of β nodes $M(u)$. We consider that $M(u)$ is *crashed* if all its nodes are crashed, which happens with probability $\lambda^\beta \leq 0.01$. Thus, if $M(u)$ is *alive*, at least one node of $M(u)$ is alive.

For two alive sets of nodes $M(u)$ and $M(v)$, let u' (resp. v') be an alive node of $M(u)$ (resp. $M(v)$). Then, after transformation 3, the channel $\{u', v'\}$ is replaced by a set of β channels. We consider that this group of channels is *crashed* if all its channels are crashed, which happens with probability $\lambda^\beta \leq 0.01$. Otherwise, u' and v' are connected by at least one channel.

Let u and v be the two nodes of Z_m such that $p \in M(u)$ and $q \in M(v)$. Then, the communication probability of p and q in W_m is at least the communication probability of u and v in Z_m when $\lambda \leq 0.01$. Thus, the situation is equivalent to the previous case, and p and q are reliably connected. □

Theorem 2. *There exists a constant Δ such that, $\forall m \geq 2$, the maximal degree of W_m is at most Δ.*

Proof. Let $m \geq 2$. The maximal degree of T_m and R_m is 3. Thus, the maximal degree of X_m is at most 9, and the maximal degree of Y_m is at most 18. After the 3 final transformations of 3.1, the maximal degree of W_m is at most $\Delta = 18\alpha\beta^2$. Thus, the result, as α and β are independent from m. □

Theorem 3. *There exists a constant C such that, $\forall m \geq 2$, the number of nodes of W_m is at most Cm.*

Proof. Let $m \geq 2$. As T_m, R_m and X_m are 3 floor graphs of height H_m, let $T_m = (V_1, \dots, V_{H_m}, E)$, $R_m = (V'_1, \dots, V'_{H_m}, E)$ and $X_m = (V^*_1, \dots, V^*_{H_m}, E)$.

According to the definition of T_m, $\forall i \in \{1, \dots, h_m\}$, $|V_i| \leq 2^{h_m - i}$, and $\forall i \in \{h_m + 1, \dots, H_m\}$, $|V_i| = 1$. According to the definition of R_m, starting from the first floor, $|V'_i|$ at most doubles every 2 floors. This is also true if we start from the last floor. Thus, $\forall i \in \{1, \dots, H_m\}$, $|V'_i| \leq 2^{i/2}$ and $|V'_i| \leq 2^{(H_m - i)/2}$.

Thus, $\forall i \in \{1, \dots, h_m\}$, $|V^*_i| = |V_i||V'_i| \leq 2^{h_m - i} 2^{i/2} = 2^{h_m - (i/2)}$, and $\forall i \in \{h_m + 1, \dots, H_m\}$, $|V^*_i| = |V_i||V'_i| \leq 2^{(H_m - i)/2}$. Thus, X_m contains at most $D = A + B$ nodes, with $A = \Sigma_{i=1}^{i=H_m} 2^{h_m - (i/2)}$ and $B = \Sigma_{i=1}^{i=H_m} 2^{(H_m/2) - (i/2)}$.

$A \leq 2\Sigma_{i=0}^{i=H_m} 2^{h_m - i} \leq 2(a + a/2 + a/4 + \dots) \leq 4a$, with $a = 2^{h_m}$. Thus, $A \leq 2^{h_m + 2}$. $B \leq 2\Sigma_{i=0}^{i=H_m} 2^{(H_m/2) - i} \leq 2(b + b/2 + b/4 + \dots) \leq 4b$, with $b = 2^{H_m/2}$. Thus, as $h_m \geq H_m/2$, $b \leq 2^{h_m}$ and $B \leq 2^{h_m + 2}$. Therefore, $D \leq 2^{h_m + 3}$.

As h_m is the smallest integer such that $2^{h_m - 1} \geq m$, we have $h_m \leq 2 + \log m$ and $D \leq 2^{5 + \log m} = 2^5 m$. Therefore, the graph Y_m contains at most $2^6 m$ nodes, and the graph W_m contains at most Cm nodes, with $C = 2^6 \alpha\beta$. Thus, the result. □

4 Solving the RBD problem

In this section, we define a graph G_n (4.1) and prove that it solves the RBD problem (4.2).

4.1 A RBD graph

Overview. For the motivation of the construction steps, please refer to the introduction. Here, we explain these steps in more details.

Let W_m be the WRBD graph defined in Section 3. Then, $\forall n \geq 2$, we consider the largest m such that the number of nodes of W_m is at most n. If such a m does not exist, we define G_n as a complete graph with redundancy of channels. As it only happens for bounded values of n, it does not break the "Bounded degree" property.

Otherwise, we consider a set V of n nodes, and we split V into subsets of $\lfloor m/2 \rfloor$ nodes. Then, we connect each pair of subsets with an instance of W_m merged with the nodes of V. The resulting graph is G_n. Doing so ensures that any two nodes of V are reliably connected. Besides, according to the "Linear number of nodes" property of W_m, the number of instances of W_m is bounded, and so is the maximal degree of G_n.

Construction of G_n. Let $n \geq 2$, and let V be a set of n nodes.

Let W_m be the graph defined in Section 3. Let N_m be the total number of nodes of W_m ($N_m \geq m$), and let S_m be the set of m nodes reliably connected by W_m.

If there exists no $m \geq 2$ such that $N_m \leq n$, then for any two nodes p and q of V, we add $\lceil \log(1-\mu)/\log(1-\lambda) \rceil$ channels between p and q ("complete graph" case).

Otherwise, let $m \geq 2$ be the largest integer such that $N_m \leq n$. Let M be the smallest integer such that $M\lfloor m/2 \rfloor \geq n$. Let $\{A_1, \ldots, A_M\}$ be a set of M subsets of V such that $\bigcup_{i \in \{1,\ldots,M\}} A_i = V$ and $\forall i \in \{1, \ldots, M\}$, $|A_i| = \lfloor m/2 \rfloor$.

Then, $\forall (i,j) \in \{1, \ldots, M\}^2$, we apply the following transformations. Let $W(i,j)$ be an instance of W_m, let $V(i,j)$ be the set of nodes of $W(i,j)$, and let $S(i,j)$ be the set of m nodes corresponding to S_m. Let $A(i,j)$ and $B(i,j)$ be two disjoint subsets of $S(i,j)$ such that $|A(i,j)| = |B(i,j)| = \lfloor m/2 \rfloor$. We merge the $\lfloor m/2 \rfloor$ nodes of $A(i,j)$ (resp. $B(i,j)$) with the $\lfloor m/2 \rfloor$ nodes of A_i (resp. A_j). Then, we merge the $N_m - 2\lfloor m/2 \rfloor$ nodes of $V(i,j) - A(i,j) - B(i,j)$ with any $N_m - 2\lfloor m/2 \rfloor$ nodes of $V - A_i - A_j$. The graph thus obtained is G_n.

4.2 Correctness

We prove that the graph G_n solves the RBD problem.

In Theorem 4, we prove the **Reliability** property. Let p and q be two nodes of G_n. In the "complete graph" case, the reliability property is ensured by the number of channels between p and q. Otherwise, it is ensured by the fact that p and q belong to the set S_m of at least one instance of W_m.

In Theorem 5, we prove the **Bounded degree** property. We first notice that the "complete graph" case only occurs when $n \leq N_2$. Thus, in this case, the degree is bounded. Otherwise, we show that the number of subsets of $\lfloor m/2 \rfloor$ nodes is bounded (which is a consequence of the linearity property of the WRBD problem). Thus, the number of instances of W_m is bounded, and so is the degree of G_n.

Theorem 4. *$\forall n \geq 2$, for any two nodes p and q of G_n, p and q are reliably connected.*

Proof. If there exists no $m \geq 2$ such that $N_m \leq n$, then p and q are connected by $k = \lceil \log(1-\mu)/\log(1-\lambda) \rceil$ channels. Thus, the probability that p and q are connected is $1-(1-\lambda)^k$. As $k \geq \log(1-\mu)/\log(1-\lambda)$, $\log(1-\mu) \geq k\log(1-\lambda)$ (as $\log(1-\lambda) < 0$). Then, $1-\mu \geq (1-\lambda)^k$, and $1-(1-\lambda)^k \geq \mu$. Therefore, p and q are reliably connected.

Otherwise, let i and j be such that $p \in A_i$ and $q \in A_j$. Then, p and q belong to the set of nodes $S(i,j)$ of the graph $W(i,j)$. Thus, according to the reliability property of the WRBD problem, p and q are reliably connected. □

Theorem 5. *There exists a constant Δ such that, $\forall n \geq 2$, the maximal degree of G_n is at most Δ.*

Proof. As the graph W_m solves the WRBD problem, there exists two constants Δ_0 and C_0 such that, $\forall m \geq 2$, the maximal degree of W_m is at most Δ_0 ("Bounded degree" property) and $N_m \leq C_0 m$ ("Linear number of nodes" property).

Let $n \geq 2$. If there exists no $m \geq 2$ such that $N_m \leq n$, then $\forall m \geq 2$, $N_m > n$. In particular, $n < N_2$. Thus, each node of S is connected to at most $\Delta_1 = N_2\lceil \log(1-\mu)/\log(1-\lambda)\rceil$ neighbors. Thus, the result, if we take $\Delta = \Delta_1$.

Otherwise, let $m \geq 2$ be the largest integer such that $N_m \leq n$. Thus, $N_{m+1} > n$, and as $N_{m+1} \leq C_0(m+1)$, $n < C_0(m+1)$. As M is the smallest integer such that $M\lfloor m/2 \rfloor \geq n$, we have $(M-1)\lfloor m/2 \rfloor < n$. Thus, $M < 1+n/\lfloor m/2 \rfloor < 1+C_0(m+1)/\lfloor m/2 \rfloor$. Then, as $(m+1)/\lfloor m/2 \rfloor \leq 4$, $M \leq 1+4C_0$.

$\forall (i,j) \in \{1,\ldots,M\}^2$, each node of V is merged with at most 2 nodes of $W(i,j)$. As the maximal degree of $W(i,j)$ is at most Δ_0, the maximal degree of G_n is at most $2\Delta_0 M^2 \leq 2\Delta_0(1+4C_0)^2$. Thus, the result, if we take $\Delta = 2\Delta_0(1+4C_0)^2$. □

5 Diameter

Our graph G_n, solving the RBD problem, turns out to have an optimal diameter (i.e., logarithmic in n). Remember that the diameter of a network corresponds to the maximal number of hops that a message has to cross, which directly impacts the communication delays.

Theorem 6 states that the diameter of a graph G_n solving the RBD problem cannot be better than logarithmic in n. Then, in Theorem 7, we state that our graph G_n has a logarithmic (and thus, optimal) diameter.

Theorem 6. *Let G_n be a graph solving the RBD problem. Then, the diameter of G_n is $\Omega(\log n)$ (i.e., at least logarithmic in n).*

Proof. As G_n solves the RBD problem, the degree of G_n can be bounded by a constant $\Delta \geq 2$. Let p be any node of G_n. Then, at most Δ nodes are at distance 1 from p, at most Δ^2 nodes are at distance 2 from p, ..., at most Δ^k nodes are at distance k from p. Thus, if D is the diameter of G_n, then G_n contains at most $1+\Delta+\Delta^2+\cdots+\Delta^D \leq 2\Delta^D$ nodes (as $\Delta \geq 2$). Thus, $n \leq 2\Delta^D$ and $D \geq (\log n - \log 2)/\log \Delta = \Omega(\log n)$. □

Theorem 7. *The graph G_n presented in this paper has a $O(\log n)$ diameter.*

Proof. Let W_m be the WRBD graph defined in Section 3. As W_m is a floor graph of height H_m, the diameter of W_m is at most $D = 2H_m$. As K_m is the smallest integer such that $2+4K_m \geq h_m$, $2+4(K_m-1) < h_m$ and $H_m = 2+4K_m < h_m+4$. As h_m is the smallest integer such that $2^{h_m-1} \geq m$, $2^{h_m-2} < m$ and $h_m < \log m + 2$. Thus, $D = 2H_m < 2(\log m + 6) = O(\log m)$.

Now, let G_n be the RBD graph defined in Section 4. As G_n is the combination of several graphs W_m of diameter $O(\log m)$ with $m \leq n$, the diameter of G_n is also $O(\log n)$. □

6 Byzantine failures

Until now, we considered *crash* failures, where the failed components (nodes and channels) simply stop functionning. When Byzantine failures are considered [17], the graph we considered so far reveals insufficient. Indeed, even one single Byzantine failure, if not contained, can potentially broadcast false messages to any other node, and deceive the whole network.

A classical strategy to contain Byzantine failures is to perform majority votes [8, 18]: a message is accepted and forwarded only if it is received throught a majority of channels. Thus, assuming there is a majority of correct components, the effect of Byzantine components can be masked by the vote. In the following, we explain how our solution can tolerate Byzantine failures by increasing the level of redundancy and adding several layers of majority votes. Yet, the main ideas remain the same.

Whilst the solution we presented (assuming only crashes) works for any failure rate $\lambda \in]0,1[$, in order to tolerate Byzantine failures, we assume $\lambda \in]0,0.5[$. This is necessary because of the classical argument of *indistinguishability* (e.g., [8] and [18]). Indeed, if a solution existed for $\lambda = 0.5$, then with the same probability, correct and Byzantine components could be exchanged. As the correct components can ensure safe communication with probability μ, the Byzantine components also could, a contraditicion. If $\lambda > 0.5$, then the Byzantine components can simulate the case $\lambda = 0.5$ by acting as correct components with probability $\lambda - 0.5$.

Now, assuming that $\lambda \in]0,0.5[$, our solution can be modified as follows to handle Byzantine failures. First, the construction scheme of the fractal graph described in Figure 4 should contain three instances of Q_k instead of two, with a majority vote at the junction. Then, the result of Lemma 1 remains correct provided that $\lambda < 0.001$[3], and the number of nodes remains linear[4]. Second, the three last transformations of 3.1 should be adapted to Byzantine failures, by increasing the level of redundancy and adding majority votes:

1. In Transformation 1 (Network replication), the number of replications α should be large enough so that the probability to have a strict majority of correct instances of Y_m is at least μ. Then, a majority vote should be performed by each node of the first floor.
2. In Transformation 2 (Node replication), the number of replications β should be large enough so that the probability to have a strict majority of correct nodes is at least 0.999 (according to the hypothesis $\lambda \leq 0.001$ above). Then, a majority vote should be performed by each node over each set of β neighbors.
3. Similarly, in Transformation 3 (Channel replication), the same number β of replications should be used. Then, a majority vote should be performed by each node over each set of β channels.

These modifications only impact the construction of the WRBD graph W_m. The construction technique of the RBD graph G_n (containing several instances of W_m) remains the same.

7 Concluding remarks

It is frequent to study the asymptotic behavior of a distributed system as a function of its number of nodes n. The parameters studied are typically the message complexity and the memory complexity. Here, we considered for the first time the asymptotic *reliability* of the network (i.e., the probability that any two nodes remain connected). We showed that it is possible to connect an arbitrarily large number of nodes with any desired level of reliability while preserving a bounded degree.

It turns out that even this apparently simple problem (with only two requirements: reliability and bounded degree) requires a non-trivial solution. Most works so far in distributed computing have focused on tolerating a specific number of failures, but a constant failure rate brings out different problems when the size of the network is unbounded (e.g., even a very small failure rate can entirely change asymptotic properties). At first glance, the desired properties may have some similarities with expander graphs [13, 10, 16]. However, these graphs are not suited for proving the reliability property: as a network is not a continuum, the combinatorial complexity of the problem explodes with the size of the network, making any proof by induction impracticable.

[3] In the proof of Lemma 1, we consider the probability that at least one instance of Q_k (out of two) is correct. Here, we should consider the probability that at least two instances of Q_k (out of three) are correct. Therefore, the formula $p(x)$ bounding the reliability becomes $(1-\lambda)^{12}(x^3 + 3x^2(1-x))$. If we assume that $\lambda \leq 0.001$, $p(x)$ keeps the same property on the interval $[0.8, 1]$, and the result of Lemma 1 remains correct.

[4] After this modification, the number of nodes of X_m is now multiplied by 4/3 every two floors (instead of 1/2). But it is still at least divided by 2 at regular intervals (every 6 floors). Thus, the argument used in Theorem 3 (i.e., $1 + 1/2 + 1/4 + 1/8 + \cdots \leq 2$) remains applicable.

Our approach suggests several research directions. For instance, an additional property could be to preserve a bounded flow of messages through each channel. One could also consider the complexity of "physically wiring" the network, and try to bound it.

References

[1] http://bluebrain.epfl.ch/page-58110-en.html.

[2] http://home.cern/about/computing.

[3] Sebastian Anthony. Microsoft now has one million servers – less than Google, but more than Amazon, says Ballmer. http://tinyurl.com/microsoft-now-has-one-million, 2013.

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. A view of cloud computing. *Communications of the ACM*, 53:50–58, 2010.

[5] Luiz André Barroso, Jimmy Clidaras, and Urs Hölzle. The datacenter as a computer: An introduction to the design of warehouse-scale machines, second edition. *Synthesis Lectures on Computer Architecture*, 2013.

[6] Paolo Costa, Austin Donnelly, Greg O'Shea, and Antony Rowstron. CamCubeOS: a key-based network stack for 3D torus cluster topologies. *22nd international symposium on High-performance parallel and distributed computing (HPDC 2013).*

[7] Csernai, Ciucu, Florin, Braun, and Gulyas. Towards 48-fold cabling complexity reduction in large flattened butterfly networks. *IEEE Conference on Computer Communications (INFOCOM 2015).*

[8] D. Dolev. The Byzantine generals strike again. *Journal of Algorithms*, 3(1):14–30, 1982.

[9] Jose Duato, Sudhakar Yalamanchili, and Ni Lionel. Interconnection networks: An engineering approach. *Morgan Kaufmann Publishers*, 2002.

[10] David Gillman. A chernoff bound for random walks on expander graphs. *SIAM Journal on Computing*, 27(4):1203–1220, 1998.

[11] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: a scalable and flexible data center network. *ACM SIGCOMM 2009 conference on Data communication*, pages 51–62.

[12] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu. DCell: a scalable and fault-tolerant network structure for data centers. *ACM SIGCOMM 2008 conference on Data communication*, pages 75–86.

[13] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

[14] Pankaj Jalote. *Fault tolerance in distributed systems.* Prentice-Hall, Inc., 1994.

[15] J. Kim, IL Evanston, W.J. Dally, S. Scott, and D. Abts. Technology-driven, highly-scalable dragonfly topology. *35th International Symposium on Computer Architecture (ISCA 2008).*

[16] Jon Kleinberg and Ronitt Rubinfeld. Short paths in expander graphs. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pages 86–95. IEEE, 1996.

[17] Leslie Lamport, Robert E. Shostak, and Marshall C. Pease. The byzantine generals problem. *ACM Trans. Program. Lang. Syst.*, 4(3):382–401, 1982.

[18] Mikhail Nesterenko and Sébastien Tixeuil. Discovering network topology in the presence of Byzantine faults. *IEEE Transactions on Parallel and Distributed Systems (TPDS)*, 20(12):1777–1789, December 2009.

[19] L. M. Ni and P. K. McKinley. A survey of wormhole routing techniques in direct networks. *Computer Journal, IEEE Computer Society Press*, 26:63–76, 1993.

[20] R. Niranjan, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat. Portland: a scalable fault-tolerant layer 2 data center network fabric. *ACM SIGCOMM 2009 conference on Data communication*, pages 39–50.

[21] Richard D Schlichting and Fred B Schneider. Fail-stop processors: an approach to designing fault-tolerant computing systems. *ACM Transactions on Computer Systems (TOCS)*, 1(3):222–238, 1983.

[22] J. Snyder. Microsoft: Datacenter Growth Defies Moore's Law. http://tinyurl.com/defy-moore-law, 2007.