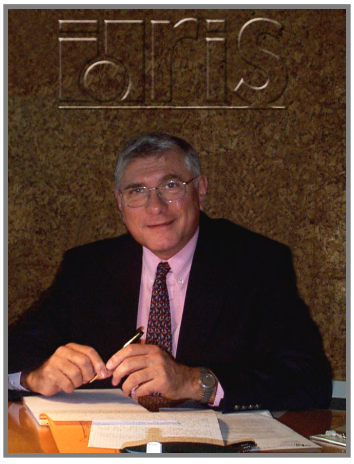
L'IDRIS à l'aube du 21e siècle

TECHNOLOGIE

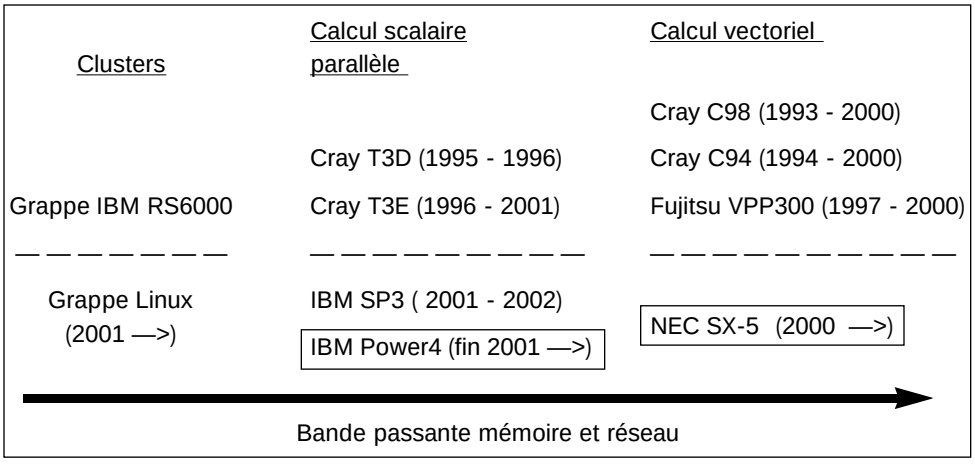
Victor Alessandrini



Depuis sa création en novembre 1993, l'IDRIS gère un environnement de calcul intensif d'avant-garde, diversifié, polyvalent et fortement évolutif.

Depuis sa création en novembre 1993, l'IDRIS a géré un environnement de calcul intensif d'avant-garde, diversifié, polyvalent et fortement évolutif. Cet environnement répond à la nécessité de contribuer, de la manière la plus efficace possible, à l'excellence de la recherche scientifique tributaire de moyens informatiques extrêmes. L'informatique scientifique est, pour l'IDRIS, un outil au service de la simulation numérique et l'essentiel de notre effort s'applique aux technologies susceptibles d'avoir un impact majeur sur l'ensemble de la recherche nationale.

La figure ci-dessous montre l'historique de l'évolution de nos moyens informatiques. La manière la plus naturelle de classifier nos supercalculateurs se base sur le niveau de raffinement et de performance des transferts de données, à l'intérieur de l'architecture et en particulier, sur la vitesse avec laquelle les données à traiter sont déplacées entre la mémoire et les processeurs. Cela constitue un élément essentiel de la performance globale de l'architecture. On trouve, à partir du sommet de l'échelle, les machines vectorielles, suivies des machines scalaires parallèles et, enfin, les grappes ou *clusters* (de PC ou autres). Cette vision de nos moyens de calcul est fondamentale pour comprendre notre stratégie. La section suivante donne une idée simple des enjeux stratégiques et technologiques associés.



▶ 1 – Les enjeux technologiques actuels

L'informatique scientifique actuelle, et l'informatique tout court, sont profondément marquées par l'arrivée en force des microprocesseurs, amorcée dans les années 80 et consolidée au début de la décennie suivante. Ce fait, ainsi que la fin de la guerre froide qui a entraîné une réduction importante des dépenses de défense aux USA, a sonné le glas des technologies dédiées au calcul scientifique de très haute performance,

La très forte évolution des microprocesseurs en technologie CMOS a facilité le développement des machines générales basées sur des technologies de masse, plus faciles à amortir sur des marchés plus larges

que le marché scientifique. On utilise le mot COST (*Commodity Off the Shelf Technology*) pour se référer aux technologies dominantes actuellement. Tous les éléments matériels des supercalculateurs scalaires d'aujourd'hui sont des éléments conçus pour un usage générique dans des systèmes diversifiés.

Les performances des microprocesseurs s'envolent mais les bandes passantes de transferts des données stagnent. En technologies COST, les architectures scalaires actuelles sont déséquilibrées.

Cette démarche, accompagnée du développement du multimédia et de l'Internet, a entraîné une formidable diffusion de l'informatique à tous les niveaux de la vie quotidienne. Toutefois, le bilan est un peu plus mitigé pour le calcul scientifique de très haute performance. Nous assistons, certes, à une augmentation spectaculaire de la puissance de calcul des microprocesseurs, depuis plus de 15 ans. La diminution constante de la taille des transistors permet d'améliorer sans cesse l'intégration à grande échelle et d'augmenter à la fois le nombre de transistors dans une puce et leur fréquence de travail. Cela se traduit par une augmentation exponentielle de leur puissance de calcul, qui double tous les 18 mois. Cependant, cette affirmation spectaculaire n'est vraie qui si les données à traiter sont déjà dans les entrailles du microprocesseur. Lorsqu'il faut que celui-ci les rapatrie depuis la mémoire — ce qui est presque toujours le cas — la situation est toute autre. Car, depuis 15 ans, les performances des accès mémoire évoluent très peu et les vitesses de transfert des données entre les mémoires et les processeurs, disponibles en technologies COST, sont devenues totalement inadaptées.

Les supercalculateurs vectoriels sont seuls, aujourd'hui, à disposer d'une bande passante de transferts des données en harmonie avec la puissance de calcul des processeurs.

Les supercalculateurs vectoriels sont la seule exception à cet état de fait. Comme leur nom le laisse entendre, ils sont spécialement adaptés aux traitements des vecteurs, c'est-à-dire des longs tableaux réguliers. Le calcul vectoriel n'est en fait qu'un cas particulier du calcul parallèle. Les machines vectorielles tirent leur performance de deux éléments : des processeurs spécialisés dans le traitement des vecteurs et d'une vitesse de transfert des données entre la mémoire et les processeurs exceptionnelle, susceptible d'alimenter les processeurs aussi vite qu'ils produisent des résultats. Puisqu'une grande partie des codes scientifiques utilise des structures de données régulières, ces calculateurs ont traditionnellement exercé un impact majeur dans le calcul scientifique. Entre 50 et 60 pour cent des projets scientifiques en cours à l'IDRIS bénéficient de ce type d'architecture.

Les premiers supercalculateurs Cray C98 et C94, installés à l'IDRIS, s'appuyaient sur des technologies rares et chères dédiées au calcul scientifique : ils étaient, de toute évidence, voués à disparaître à long terme. Cela s'est passé ainsi aux USA où le calcul vectoriel est aujourd'hui pratiquement inexistant. Pourtant, les constructeurs japonais NEC et FUJITSU sont parvenus, au milieu des années 90, à donner un nouvel élan au calcul vectoriel — hors USA, bien entendu — en produisant des supercalculateurs vectoriels très puissants, en technologie CMOS. La technologie de base est la même que celle de l'informatique « grand public » mais ces calculateurs demeurent néanmoins des machines pointues et dédiées au calcul scientifique, en raison des formidables bandes passantes mémoires — processeurs mises en jeu pour garantir leur performance. De tels éléments, même en technologie CMOS, ne sont pas des fournitures courantes.

▶ 2 – La démarche stratégique de l'IDRIS

La stratégie de l'IDRIS se base sur la recherche de performances soutenues exceptionnelles et de logiciels système matures, robustes et fiables.

Le renouvellement du parc des machines, pour l'adapter aux enjeux scientifiques de la nouvelle décennie, a été effectué en 1999 et 2000, dans le contexte technologique que nous venons de décrire. Fidèle à ses objectifs, la stratégie de l'IDRIS a été la recherche de performances soutenues exceptionnelles, par l'intermédiaire d'architectures bien équilibrées, robustes et tolérantes aux pannes, dotées de logiciels système matures et fiables, capables de gérer correctement les conflits qui émanent du partage d'un instrument de recherche par des centaines, voire des milliers de scientifiques utilisateurs.

Des besoins de cette envergure demandent des machines exceptionnelles. La qualité des logiciels maîtres qui pilotent l'ordinateur et arbitrent les conflits, les bonnes performances d'entrées-sorties d'immenses masses de données, sont des besoins difficiles à satisfaire avec des architectures COST relativement précaires comme les *clusters* de PC. Ces architectures sont, certes, financièrement très abordables. Mais, ce qui compte, n'est pas seulement le coût d'une architecture, mais la qualité et la quantité des résultats scientifiques obtenus. De ce point de vue, les machines de l'IDRIS sont probablement 100 fois plus chères qu'un bon *cluster* de PC, mais elles sont plus de 100 fois plus rentables : elles traitent en effet plus de 400 projets scientifiques qui ne seraient pas traitables sur 100 *clusters* de PC...

L'autre élément stratégique majeur dans la démarche de l'IDRIS résulte de la nécessité de prévoir l'avenir dans un domaine en mutation technologique rapide. Ainsi, notre objectif majeur n'est pas la recherche d'une solution ou d'une réponse immédiate aux besoins du moment mais plutôt l'identification d'une technologie capable de tenir un rôle de vecteur de développement du calcul intensif de haute performance.

Fin 1998 et suite à une prospection très approfondie effectuée par le Conseil scientifique de l'IDRIS, le renouvellement du parc des supercalculateurs fut lancé. L'objectif était la multiplication par 10 de la puissance de calcul installée sur le site, à la fois pour le calcul scalaire et vectoriel. Le renforcement vectoriel fut jugé incontournable. La mise en place d'une architecture scalaire de nouvelle génération est alors apparue indispensable pour remplacer le Cray T3E qui commençait à dater et pour prendre en charge les 40 à 50 pour cent d'applications mal adaptées au calcul vectoriel.

L'objectif majeur de l'IDRIS est l'identification des technologies capables de tenir un rôle de vecteur de développement du calcul intensif de haute performance.

▶ 3 – L'évolution du calcul vectoriel

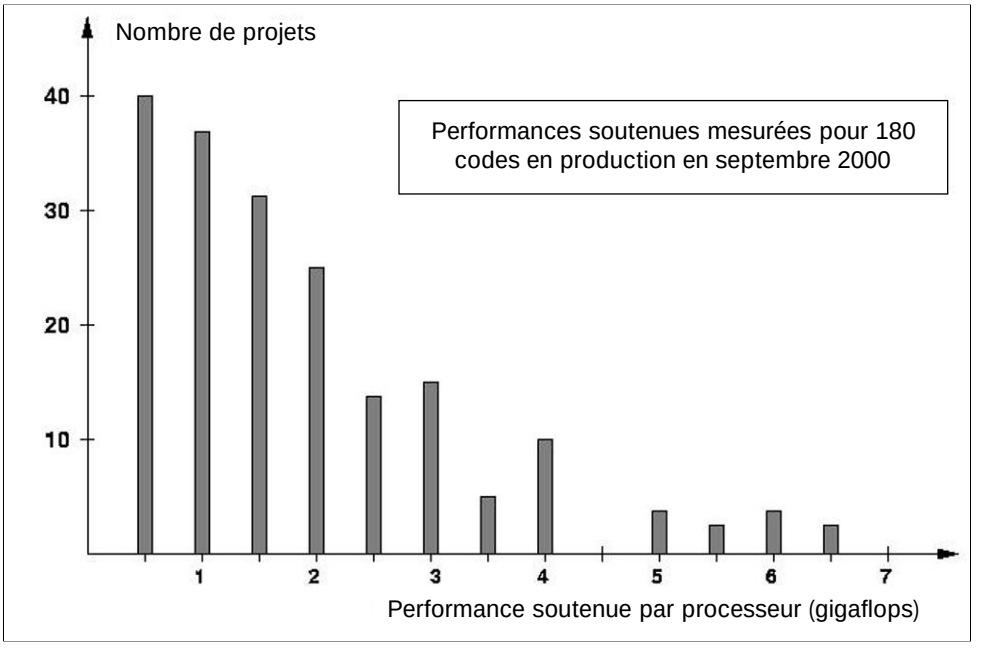
Les deux supercalculateurs vectoriels installés à la création de l'IDRIS, les Cray C98 et C94, ont rendu d'excellents services à la recherche en France jusqu'à leur arrêt en février 2000 après plusieurs années d'exploitation intensive avec un rendement exceptionnel. Ils ont été accompagnés par un supercalculateur vectoriel Fujitsu VPP300 à partir de 1997 qui a doublé la puissance de calcul et est resté en service aussi jusqu'en février 2000.

La première opération de renouvellement du parc a été le remplacement

du parc vectoriel préexistant par une grappe de trois supercalculateurs vectoriels NEC SX-5. Inaugurée en juin 2000 et comportant actuellement 40 processeurs de 8 gigaflops chacun et 236 gigaoctets de mémoire globale, cette architecture est actuellement le plus gros supercalculateur vectoriel en France et le plus gros système NEC dans le monde.

La grappe NEC SX-5 accueille en 2001 plus de 250 projets scientifiques. Son taux d'utilisation ne cesse de croître depuis sa mise en service : il est aujourd'hui de 65 pour cent. L'IDRIS suit de très près les performances réelles et soutenues de tous les codes qui s'exécutent sur la grappe SX-5. Nous présentons ci-après les performances soutenues, par processeur, d'environ 180 codes qui étaient en pleine production en septembre 2000. La figure suivante est un histogramme qui montre le nombre de codes qui s'exécutent à une performance donnée.

Le *cluster* vectoriel NEC actuellement en exploitation à l'IDRIS est le plus gros super calculateur vectoriel en France, le deuxième en Europe et le plus gros système NEC dans le monde.



Les performances soutenues de plusieurs gigaflops par processeur obtenus sur NEC SX-5, sont totalement hors de portée des machines scalaires présentes ou à venir à court et moyen termes.

Ce résultat témoigne de la formidable efficacité du calcul vectoriel, lorsqu'il s'applique à un problème adapté. Des performances soutenues de plusieurs gigaflops par processeur sont totalement hors de portée des machines scalaires présentes ou à venir à court et moyen terme. La robustesse du système d'exploitation et de l'environnement de production, la qualité des logiciels d'arbitrage des conflits et de gestion des ressources, les performances d'entrées-sorties, la diversité des modèles de programmation, la capacité d'accès à une énorme mémoire partagée, ajoutées à la profusion de gigaflops, font de cette plateforme un outil de recherche exceptionnel.

▶ 4 – L'évolution du calcul scalaire parallèle

Dès le milieu des années 90, l'IDRIS s'est lancé vigoureusement dans la direction du calcul scalaire parallèle. Il semblait à l'époque que ces architectures remplaceraient à terme les architectures vectorielles (nous avons vu que cela n'a pas encore été le cas). Un supercalculateur

Cray T3D fut installé en 1995 (128 processeurs Alpha EV4 et 16 gigaoctets de mémoire totale) et remplacé en 1996 par l'actuel Cray T3E (256 processeurs Alpha EV5 à 600 mégaflops par processeur, 32 gigaoctets de mémoire totale). Ces architectures constituent une superbe réalisation technologique, comportant un très grand nombre de fonctionnalités dédiées à l'informatique scientifique. Le T3E s'appuie, certes, sur des microprocesseurs standards, mais il incorpore, au niveau de son architecture, un nombre important d'éléments matériels spécialement conçus pour le calcul de haute performance. Il se situe au sommet d'une démarche technologique, par la suite en déclin, de par la démarche COST.

Le vecteur du calcul scalaire parallèle de très haute performance à l'IDRIS sera la nouvelle architecture IBM Power4. Car, en ce début de siècle, la démarche stratégique de la société IBM s'accorde parfaitement à la vision stratégique de l'IDRIS.

Après plusieurs années d'exploitation très intensive, cette plate-forme sera arrêtée fin 2001, après la mise en service d'un supercalculateur scalaire de nouvelle génération. Son remplacement s'annonçait difficile : les critères de performance et de qualité adoptés par l'IDRIS sont peu compatibles avec les environnements logiciels parfois précaires qui accompagnent les *clusters* en technologie COST. Heureusement, une option séduisante nous attendait : le vecteur du calcul scalaire parallèle de très haute performance à l'IDRIS, à l'aube du nouveau millénaire, sera la nouvelle architecture IBM Power4. Car, en ce début de siècle, la démarche stratégique de la société IBM s'accorde parfaitement avec la vision stratégique de l'IDRIS. Le projet Power4 implique l'abandon de la démarche COST : on arrête de concevoir et de fabriquer des microprocesseurs et des composants génériques et l'on s'attaque à une nouvelle architecture globale où tous les éléments (processeurs, communications avec la mémoire, communications entre processeurs, entrées-sorties,...) sont repensés et harmonisés. Cette démarche, très proche de celle que nous avons connue à l'époque du T3E, était de nature à nous séduire. L'IDRIS a trouvé, pour le T3E, un très digne successeur.

Dans l'architecture IBM Power4, les bandes passantes se rapprochent de celles des machines vectorielles, et le système fonctionne comme un tout cohérent et équilibré.

L'élément de base de cette nouvelle technologie est un module appelé MCM (*Multi-Chip Module*), un carré de moins de 12 cm de coté, qui est déjà un supercalculateur à part entière. Le MCM comporte, dans chaque angle du carré, une puce bi-processeur de 4,4 gigaflops de puissance de crête chacun. Cela fait donc un total de huit processeurs et environ 40 gigaflops de puissance de crête dans le MCM, plus des mémoires caches, des contrôleurs de communications et d'entrées-sorties et, surtout, des bandes passantes internes nettement plus importantes que celles disponibles actuellement en technologies COST. Chaque bi-processeurs dans un coin du MCM peut échanger plus de 30 gigaoctets par seconde avec le reste du système : les bandes passantes commencent à se rapprocher de celles des machines vectorielles. De surcroît, ce système fonctionne comme un tout cohérent et équilibré : si la fréquence de travail augmente, les bandes passantes augmentent au même rythme que les performances des processeurs. Enfin, seule la très forte chaleur dégagée interdit de mettre ce supercalculateur dans sa poche.

Quatre MCM peuvent être assemblés et interconnectés par un réseau à très haut débit, de même nature et de débit comparable à celui qui intervient à l'intérieur du MCM. On aboutit ainsi à un système à 32 processeurs, le nœud Power4 qui apparaît à l'utilisateur comme un

système multiprocesseur symétrique (SMP) à mémoire partagée. Un nœud Power4 représente donc une puissance de calcul d'environ 150 gigaflops et une mémoire partagée pouvant aller jusqu'à 256 gigaoctets. Chaque nœud apporte ainsi une puissance de calcul scalaire et une taille mémoire comparables à un gros nœud vectoriel NEC SX-5. Comme les nœuds NEC SX-5, les nœuds IBM Power4 sont des boîtes à très grande bande passante, comme l'IDRIS les affectionne. Cela promet des rendements et des puissances de calcul soutenues, supérieurs à ceux des machines scalaires actuelles.

Le mot « *constellation* » désigne une architecture en *cluster* constitué d'un nombre restreint de nœuds hyper-puissants, à grosse mémoire partagée.

Cette architecture innovante est susceptible d'ouvrir des perspectives nouvelles dans le domaine du calcul de très haute performance. Nous nous attendons à ce qu'elle rapproche le calcul scalaire du calcul vectoriel, en fournissant des puissances soutenues supérieures aux 10-15 pour cent de la performance crête caractéristique des machines scalaires d'aujourd'hui. La configuration qui sera installée à l'IDRIS fin 2001 sera constituée de 8 nœuds Power4, interconnectés par un réseau Colony. Elle aboutira à une puissance nominale de 1,2 téraflops et à une mémoire globale de 832 gigaoctets (la puissance nominale du T3E est de 150 gigaflops et sa mémoire globale de 32 gigaoctets !). Chacun des huit nœuds Power4 est plus puissant que le T3E.

► 5 – Les moyens de calcul du 21e siècle

Le mot « *constellation* » a été utilisé pour la première fois l'an dernier, à la conférence internationale d'informatique scientifique de Dallas (SC2000), pour désigner un *cluster* constitué d'un nombre restreint de nœuds hyper puissants, à grosse mémoire partagée.

A l'aube du 21e siècle, les constellations vectorielles (NEC) et scalaires (IBM) de l'IDRIS devraient contribuer efficacement à repousser les limites de nos connaissances.

L'IDRIS entame le 21e siècle avec des constellations scalaires (IBM) et vectorielles (NEC) dont les nœuds possèdent une puissance comprise entre 128 et 160 gigaflops et une taille mémoire variant de 64 à 256 gigaoctets. De surcroît, ces constellations ne seront pas totalement autonomes : des évolutions en cours dans le domaine du calcul réparti permettront de les utiliser de manière harmonieuse et complémentaire, dans un court délai.

Ces constellations sont intégrées dans un environnement complexe comprenant, d'une part, un nombre important de machines de support et de serveurs de visualisation et de pré et post-traitement performants et, d'autre part, un système d'avant-garde de gestion de données, capable de stocker jusqu'à 800 téraoctets de données produites par les simulations numériques.

Un tel environnement, accompagné des services à très forte valeur ajoutée de support à tous les niveaux et de formation fournis par l'IDRIS, constitue un atout important pour la simulation numérique. Utilisé de manière de plus en plus prospective et innovante, il est à même de stimuler fortement la créativité des chercheurs et de contribuer efficacement à repousser les limites de nos connaissances.