La virgule flottante, d'Archimède au porte-clés d'Andy Grove

Jean-Michel Muller

Café gourmand scientifique du LIP
décembre 2019

Le banquier infernal

Voulant sécuriser ma retraite, j'ai

$e - 1 = 1.71828182845904523536028747135266249775724709 3 \ldots$

euros à placer...

je me rends à la Société chaotique de banque, qui fait de la pub pour de nouveaux placements...

Le banquier infernal

À la Société chaotique de banque, le banquier m'explique :

Le banquier infernal

À la Société chaotique de banque, le banquier m'explique :

- la première année, mon capital est multiplié par 1, et on me retire 1 euro pour frais de gestion ;

Le banquier infernal

À la Société chaotique de banque, le banquier m'explique :

- la première année, mon capital est multiplié par 1, et on me retire 1 euro pour frais de gestion ;
- la deuxième année, mon capital est multiplié par 2, et on me retire 1 euro pour frais de gestion ;

Le banquier infernal

À la Société chaotique de banque, le banquier m'explique :

- la première année, mon capital est multiplié par 1, et on me retire 1 euro pour frais de gestion ;
- la deuxième année, mon capital est multiplié par 2, et on me retire 1 euro pour frais de gestion ;
- la troisième année, mon capital est multiplié par 3, et on me retire 1 euro pour frais de gestion ;
- . . .

Le banquier infernal

À la Société chaotique de banque, le banquier m'explique :

- la première année, mon capital est multiplié par 1, et on me retire 1 euro pour frais de gestion ;
- la deuxième année, mon capital est multiplié par 2, et on me retire 1 euro pour frais de gestion ;
- la troisième année, mon capital est multiplié par 3, et on me retire 1 euro pour frais de gestion ;
- . . .
- la 25ème année, mon capital est multiplié par 25, et on me retire 1 euro pour frais de gestion ;

Au bout de 25 ans, je peux retirer mon argent. . . est-ce intéressant ?

Je n'ai pas signé tout de suite...

J'ai cherché à calculer ce que serait mon capital au bout de 25 ans. . .

Je n'ai pas signé tout de suite...

J'ai cherché à calculer ce que serait mon capital au bout de 25 ans...

- ma calculette (Casio) : -747895876335 euros ;

Je n'ai pas signé tout de suite...

J'ai cherché à calculer ce que serait mon capital au bout de 25 ans...

- ma calculette (Casio) : -747895876335 euros ;
- mon ordinateur (Proc. Intel Xeon, compilateur gcc, sous Linux) : $+1201807247$ euros ;

Je n'ai pas signé tout de suite...

J'ai cherché à calculer ce que serait mon capital au bout de 25 ans...

- ma calculette (Casio) : -747895876335 euros ;
- mon ordinateur (Proc. Intel Xeon, compilateur gcc, sous Linux) : $+1201807247$ euros ;
- en fait, la « vraie » valeur est d'environ 0.0399 euros...

Double conclusion de ce fâcheux épisode

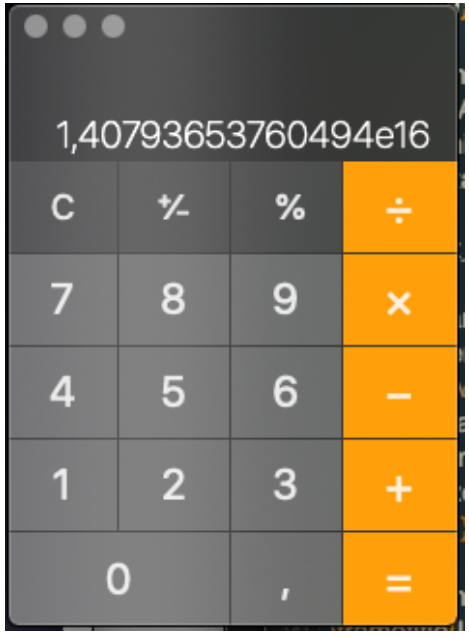
Double conclusion de ce fâcheux épisode

1. ne faites pas aveuglément confiance à votre ordinateur ;

Double conclusion de ce fâcheux épisode

1. ne faites pas aveuglément confiance à votre ordinateur ;
2. ne faites pas aveuglément confiance à votre banquier.

Notation « scientifique » de nos calculatrices



Notation « scientifique » de nos calculatrices

$1,407936537604916$ représente $1,40793653760494 \times 10^{16}$,
c'est-à-dire :

$$1,40793653760494 \times \underbrace{10 \times 10 \times 10 \times \cdots \times 10}_{16 \text{ fois}} = 14079365376049400$$

$\rightarrow$ Base 10.

Arithmétique virgule flottante

On généralise cela à la base β (qui vaut souvent 2) :

$$x = x_0.x_1x_2\cdots x_p \times \beta^{e_x}$$

(pareil : $\beta^3 = \beta \times \beta \times \beta$, et $\beta^{-3} = 1/\beta^3$).

Avantages pour le calcul :

- dynamique : représenter de très petits et de très grands nombres de manière compacte ;
- algorithmes arithmétiques simples.
  (ce qui n'est pas le cas de tous les systèmes de numération : calculez MMMDCCLXLL $\times$ MXLVIII).

Les Mésopotamiens inventent les mantisses. . .

- actuel Irak, vers -2000 ;
- Système de base 60 (58 tables de multiplication à connaître !) ;
- pas de zéro « à la fin » : on manipule juste des mantisses (comme si dans notre système 25, 0.025 et 250 avaient la même représentation).

Les Mésopotamiens inventent les mantisses. . .

- actuel Irak, vers -2000 ;
- Système de base 60 (58 tables de multiplication à connaître !) ;
- pas de zéro « à la fin » : on manipule juste des mantisses (comme si dans notre système 25, 0.025 et 250 avaient la même représentation).

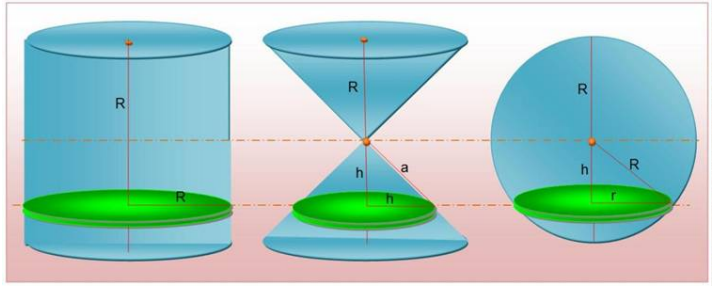
. . . et Archimède (-287 – -212) invente les exposants

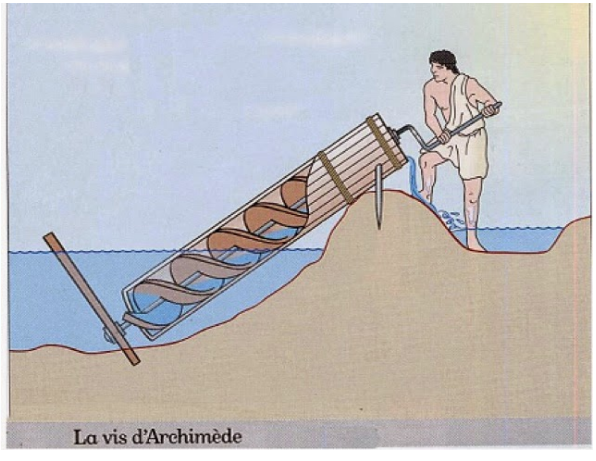
- Traité l'Arénaire (compteur de sable : *arena* = sable en Latin) ;
- nombre de grains de sable qui pourraient remplir l'Univers ;
- notation exponentielle pour représenter les ordres de grandeur.

. . . et Archimède (-287 – -212) invente les exposants

- Traité l'Arénaire (compteur de sable : *arena* = sable en Latin) ;
- nombre de grains de sable qui pourraient remplir l'Univers ;
- notation exponentielle pour représenter les ordres de grandeur.

C'est Le génie scientifique de l'antiquité.



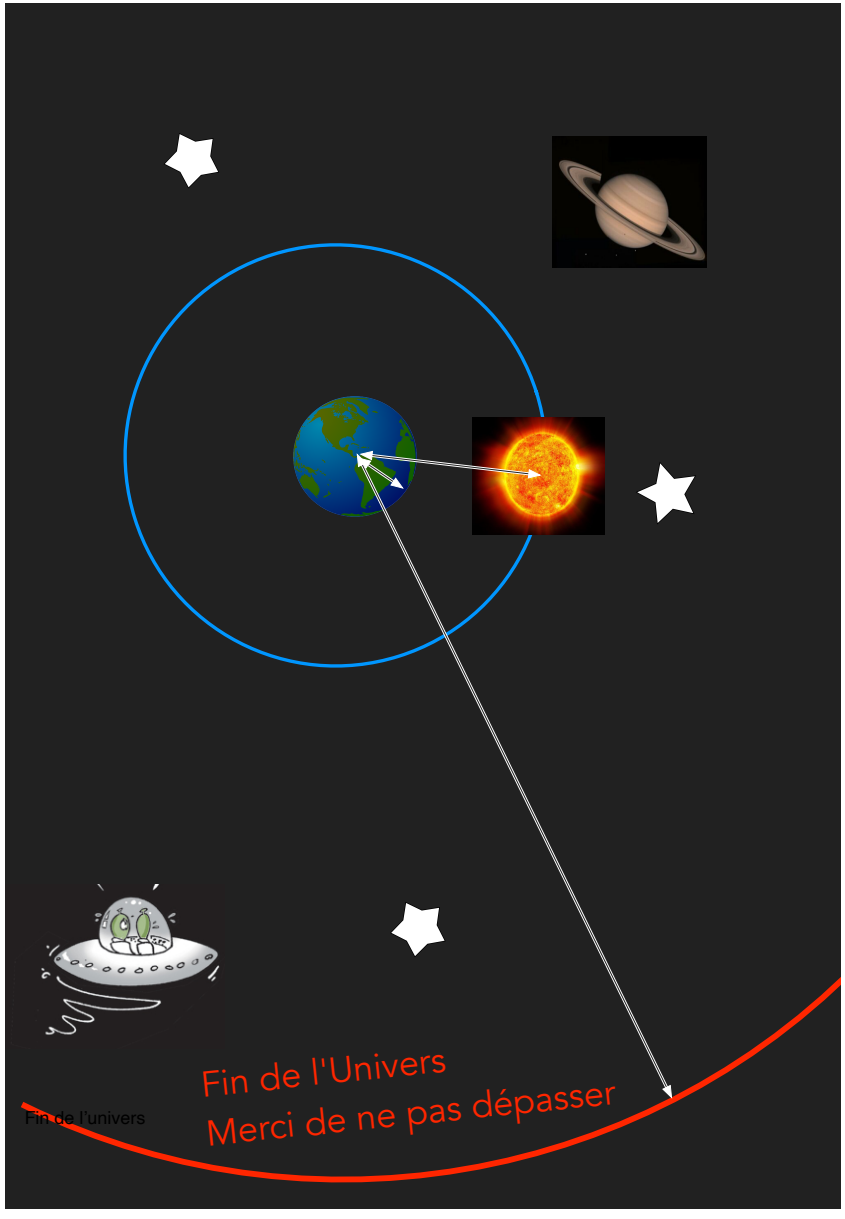


. . . et Archimède invente les exposants

Hypothèse :

$$\frac{\text{rayon Univers}}{\text{distance Terre-Soleil}}$$

$$= \frac{\text{distance Terre-Soleil}}{\text{rayon Terre}}$$



. . . et Archimède invente les exposants

- point de départ : savait compter en Grec jusqu'à 10^8 (une myriade de myriades – $\mu\upsilon\rho\iota\alpha\varsigma = 10000$) ;
- nombres de la première période :
  - nombres « premiers » : $1 \to 10^8$;
  - nombres « seconds » : de la forme $10^8 \times$ nombre « premier » ;
  - nombres « troisièmes » : de la forme $10^8 \times$ nombre « second » ;
  - . . .
  - jusqu'aux nombres « 10^8èmes » $\to \Omega = 10^{8 \cdot 10^8}$;
- nombres de la deuxième période : $\Omega \times$ nombres de la 1ère période.

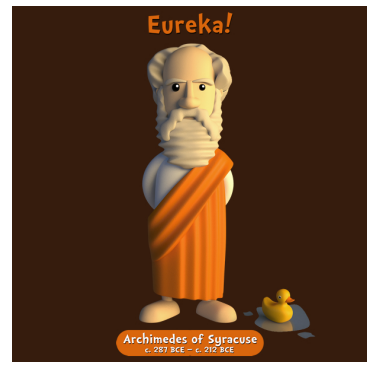
Réponse d'Archimède : on fait tenir environ 10^{63} grains de sable dans l'Univers.

. . . et Archimède invente les exposants

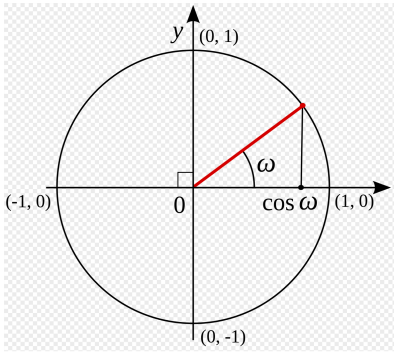
- point de départ : savait compter en Grec jusqu'à 10^8 (une myriade de myriades – $\mu\upsilon\rho\iota\alpha\varsigma = 10000$) ;
- nombres de la première période :
  - nombres « premiers » : $1 \to 10^8$;
  - nombres « seconds » : de la forme $10^8 \times$ nombre « premier » ;
  - nombres « troisièmes » : de la forme $10^8 \times$ nombre « second » ;
  - . . .
  - jusqu'aux nombres « 10^8èmes » $\to \Omega = 10^{8 \cdot 10^8}$;
- nombres de la deuxième période : $\Omega \times$ nombres de la 1ère période.

Réponse d'Archimède : on fait tenir environ 10^{63} grains de sable dans l'Univers.

Il n'a pas vraiment fait exprès mais il avait raison !



Au commencement était la trigonométrie. . .



Prostaphérèse :

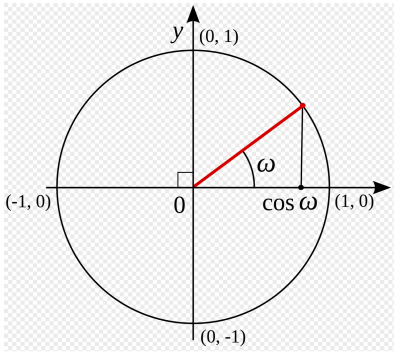
$$\cos(a)\times\cos(b) = \frac{1}{2}\Big(\cos(a+b)+\cos(a-b)\Big)$$

On réalise une bonne fois pour toutes une table des cosinus, et on pourra remplacer les multiplications par des additions !

$$A \times B \,?$$

- chercher dans la table a et b t.q. $A = \cos(a)$ et $B = \cos(b)$;
- calculer $s = a + b$ et $d = a - b$;
- chercher dans la table $S = \cos(s)$ et $D = \cos(d)$;
- le résultat est $\frac{1}{2}(S + D)$.

Au commencement était la trigonométrie...



Prostaphérèse :

$$\cos(a)\times\cos(b) = \frac{1}{2}\Big(\cos(a+b)+\cos(a-b)\Big)$$

On réalise une bonne fois pour toutes une table des cosinus, et on pourra remplacer les multiplications par des additions !

$A \times B$?

- chercher dans la table a et b t.q. $A = \cos(a)$ et $B = \cos(b)$;
- calculer $s = a + b$ et $d = a - b$;
- chercher dans la table $S = \cos(s)$ et $D = \cos(d)$;
- le résultat est $\frac{1}{2}(S + D)$.

→ 4 lectures de table et 3 additions/soustractions : faut vraiment pas aimer les multiplications.

→ Méthode connue de Ibn Jûnus au Xème siècle.

Les logarithmes puis la règle à calculs

- Neper (1650–1617) : moyen plus simple de transformer les multiplications en additions

$$\log(a \times b) = \log(a) + \log(b)$$

- Briggs : 1ères tables de logarithmes en 1617 ;
- Gunter (1624) – 10 ans après l'invention des logarithmes. Echelle fixe : distances reportées à l'aide d'un compas.

- Règles glissantes : Wingate (1627) ;

La «notation scientifique» des nombres réels

Première étape : notation a^n pour $a \times a \times \cdots \times a$ – Descartes, dans *La Géométrie* (il y invente aussi le symbole $\sqrt{\cdot}$). 1637 ?

gnes ſur le papier, & il ſuffiſt de les deſigner par quelques lettres, chaſcune par vne ſeule. Comme pour adiouſter la ligne B D a G H, ie nomme l'vne a & l'autre b, & eſcris $a + b$; Et $a - b$, pour ſouſtraire b d'a; Et ab, pour les multiplier l'vne par l'autre; Et $\frac{a}{b}$, pour diuiſer a par b; Et aa, ou a^2, pour multiplier a par ſoy meſme; Et a^3, pour le multiplier encore vne fois par a, & ainſi a l'infini; Et $\sqrt{a^2 + b^2}$, pour tirer la racine quarrée d'$a^2 + b^2$; Et $\sqrt{C.\ a^3 - b^3 + abb}$, pour tirer la racine cubique d'$a^3 - b^3 + abb$, & ainſi des autres.

vſer de chiffres en Geometrie.

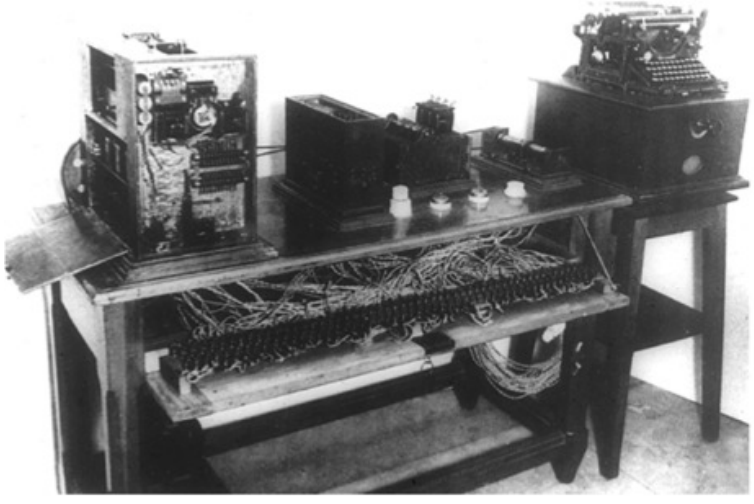
La «notation scientifique» des nombres réels

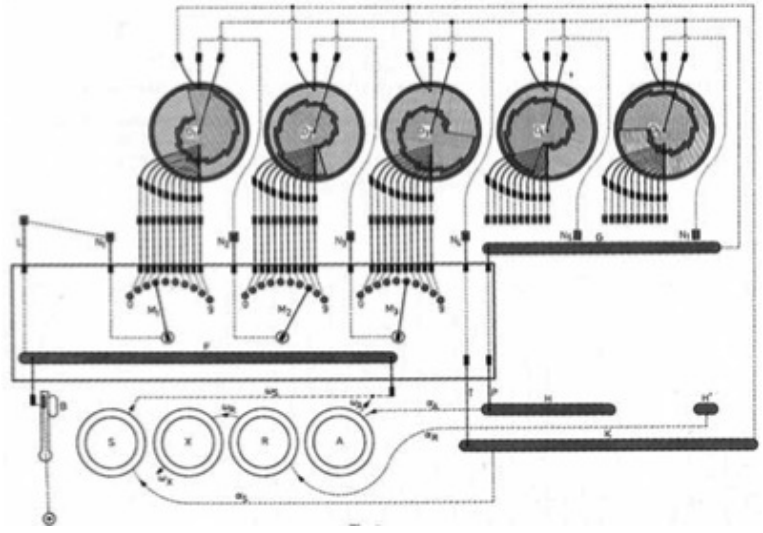
- à cause de ceci on attribue parfois l'invention de la « notation scientifique » à Descartes ;
- Wallis (1665) puis Newton (1669) : exposants négatifs, rationnels ;
- la notation x^n permet d'écrire un nombre sous la forme $m \times 10^e$, mais cette représentation ne se généralise vraiment qu'au 19ème siècle.

Internet est amusant : sur un site américain la notation scientifique a été *inventée par Descartes puis améliorée par Archimède.*

Leonardo Torres y Quevedo (1852–1936)

- version électromécanique (à relais) de la machine analytique de Babbage ;
- première proposition d'une arithmétique virgule flottante (1914) ;
- Arithmomètre (1920), opérateurs arithmétiques.



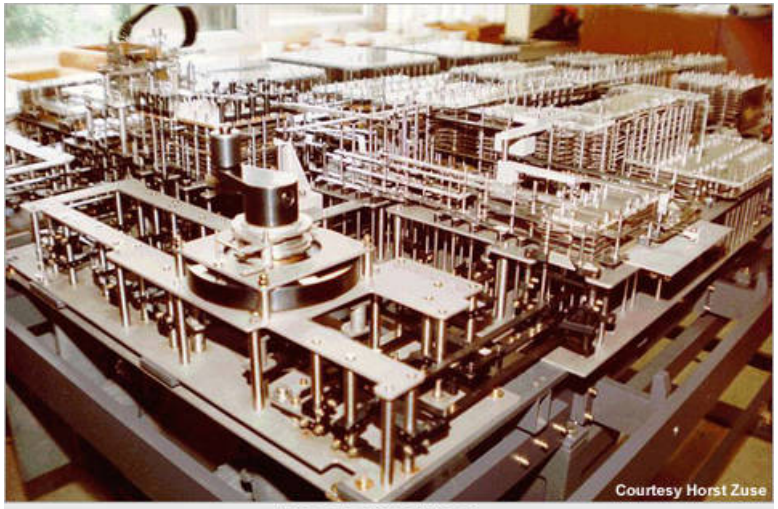


Téléphérique des chutes du Niagara (1916).

Telekino : 1ère (ou 2ème) machine radio-commandée (1901).

Konrad Zuse (1910–1995)

- Z1 (1936–1938) : calculateur mécanique ;
- Z2 (1938) : relais électromécaniques (relais de téléphone d'occasion) et mémoire mécanique ;

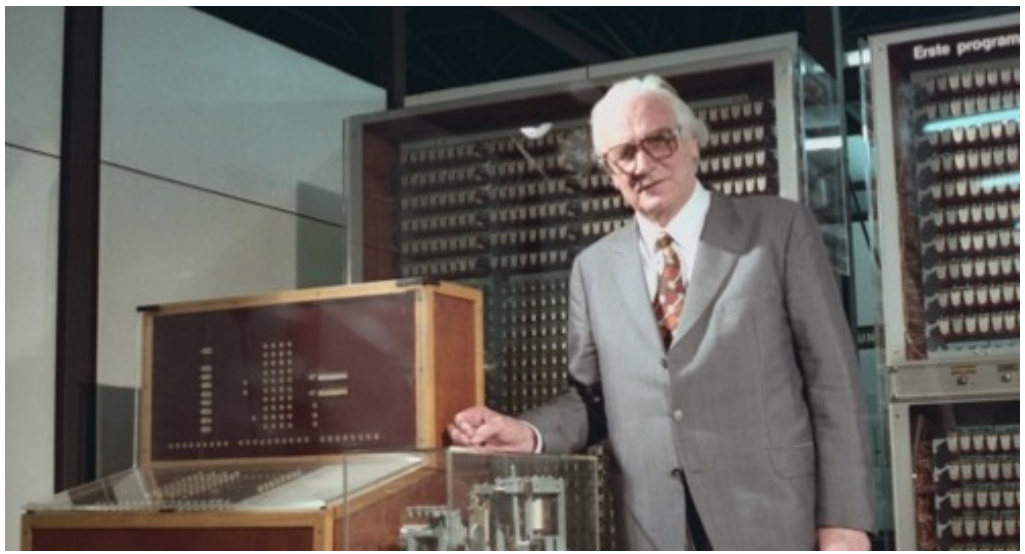


Reconstruction of the Z1

- le Z2 n'était pas fiable mais a suffi comme « preuve de concept » :
  - à convaincre Zuse qu'un calculateur d'envergure était réalisable ;
  - à convaincre le DVL (institut allemand d'aéronautique) de financer ses travaux.

Le Z3 (1941)

- arithmétique virgule flottante ;
- base 2, nombres sur 22 bits (chiffres binaires) :
  - mantisses de 14 bits ;
  - exposants de 7 bits ;
  - 1 bit de signe ;
- représentations spéciales pour $\pm\infty$ et résultats indéterminés ;
- contrairement aux Z1 et Z2, a été complètement opérationnel ;
- Zuse ne l'a pas conçu dans cette optique, mais le Z3 était un calculateur universel.



Zuse posant devant une reconstruction du Z3

Quelques autres réalisations de Zuse

- premier langage de programmation de «haut niveau» : le Plankalkül (1942–1946) ;

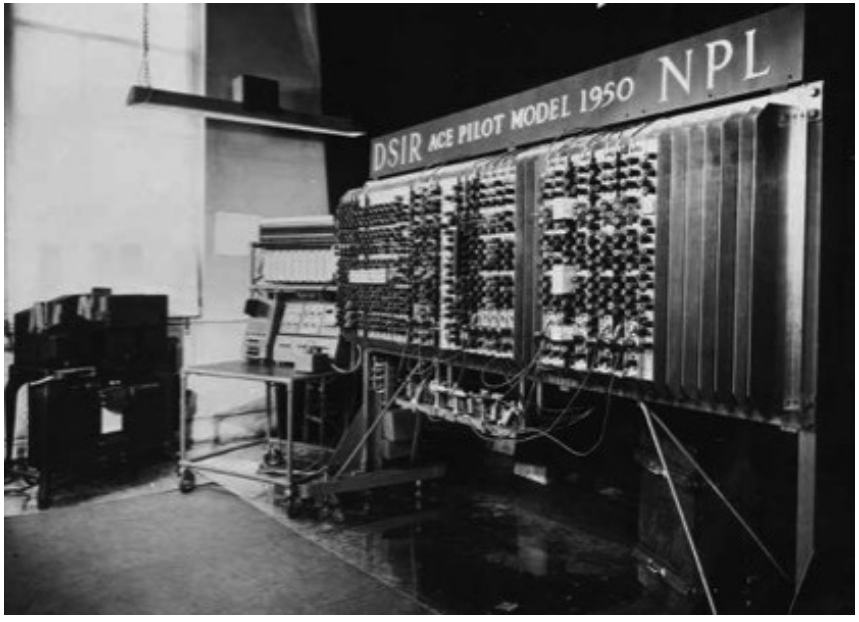
```
P1 max3 (V0[:8.0], V1[:8.0], V2[:8.0]) => R0[:8.0]
max(V0[:8.0], V1[:8.0]) => Z1[:8.0]
max(Z1[:8.0], V2[:8.0]) => R0[:8.0]
END
P2 max (V0[:8.0], V1[:8.0]) => R0[:8.0]
V0[:8.0] => Z1[:8.0]
(Z1[:8.0] < V1[:8.0]) ? V1[:8.0] => Z1[:8.0]
Z1[:8.0] => R0[:8.0]
END
```

  L'université libre de Berlin a écrit un compilateur en 2000 ;

- Calculateurs S1 et S2 : aérodynamique de bombes à guidage (précurseurs des V1). Probablement récupérés par l'URSS en 1945.

Zuse était aussi un peintre

Une machine de Turing ni en papier ni en Lego



Le *Pilot ACE*, dont le 1er programme a tourné en mai 1950

- ACE : *Automatic Computing Engine. . .* l'autre machine de Turing ;
- National Physics Laboratory, 1946 ;
- Turing quitte le projet en 1947, Wilkinson en prend le contrôle ;
- programmes VF et programmes *multi-précision* écrits par Alway et W. en 1947,avant même que le Pilot ACE ne fonctionne.

Ensuite c'est le bordel...

- Base : 2, 4, 8, 10, 16, pas la même manière de gérer $1/0$, $0/0$, $\sqrt{-1}$, etc. ;
- spécification floue des opérations ;
- Quand seule la vitesse compte : sur les Crays, le dépassement de capacité était calculé à partir des exposants des entrées, en parallèle avec le calcul effectif du produit

  → `1 * x` peut faire un overflow ;
- sur les mêmes, seuls 12 bits de x étaient examinés pour détecter une division par 0 lors du calcul y/x

  → `if (x = 0) then z := 17.0 else z := y/x`
  peut provoquer une erreur « division par zéro »...
  mais comme le multiplieur aussi ne regarde que 12 bits pour décider qu'une opérande est nulle,

  ```
  if (1.0 * x = 0) then z := 17.0 else z := y/x
  ```

  ne pose plus de problème.

Standard IEEE 754-1985

- sous l'impulsion de W. Kahan (Prof. Berkeley) ;
- choix de la base 2, de formats (32 bits, 64 bits) ;
- deux idées fortes :
  - système clos : même les opérations « illicites » $(1/0\,;\,\sqrt{-5})$ fournissent un résultat, qui doit pouvoir être réutilisable en entrée ;
  - arrondi correct : une fonction d'arrondi $\circ$ étant choisie, le calcul en machine de $a \star b$ donne

$$\circ(a \star b)$$

$\rightarrow$ amélioration de la *portabilité*, de la *prouvabilité* et de la *qualité numérique* des programmes.

Erreur de l'addition (Dekker)

Théorème 1 (Fast2Sum (Dekker))

(base ≤ 3) Soient a et b des nombres VF vérifiant $|a| \geq |b|$. Algorithme suivant : s et r t.q.

- $s + r = a + b$ *exactement ;*
- s *est « le » nombre VF le plus proche de* $a + b$.

Programme C 1

```
s = a+b;
z = s-a;
r = b-z;
```

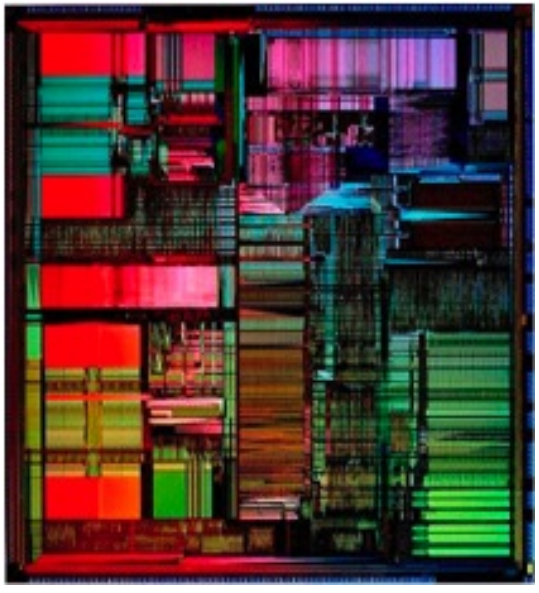
Se méfier des compilateurs « optimisants ».

Arithmétique (presque) bien spécifiée

Tout est prêt pour faire des preuves rigoureuses sauf qu'à l'époque...

- les ingénieurs/scientifiques n'en éprouvent pas vraiment le besoin : ils font de la simulation tranquilles au sol ;
- pour prouver un algorithme, il faut le connaître : culte du secret ;
- il n'y avait pas encore eu de très gros problème ;
- ...et puis chez Intel, Motorola, etc. il y avait à ce moment là un côté « bidouilleur » sympathique mais dangereux.

Automne 1994 : la précision d'une règle à calcul



- Thomas Nicely (Lynchburg Univ.) : constante de Brun

$$\left(\frac{1}{3}+\frac{1}{5}\right)+\left(\frac{1}{5}+\frac{1}{7}\right)+\left(\frac{1}{11}+\frac{1}{13}\right)+\cdots$$

(couples de nombres 1ers jumeaux). Viggo Brun, 1919 : la série converge.

- résultats pas en accord avec les précédents. Dans un tel cas on soupçonne :
  1. le programme ;
  2. le compilateur ;
  3. en dernier recours le processeur.
- le Pentium donnait un résultat incorrect pour 1/824633702441 (824633702441 et 824633702443 sont jumeaux).

Le « bug » du Pentium

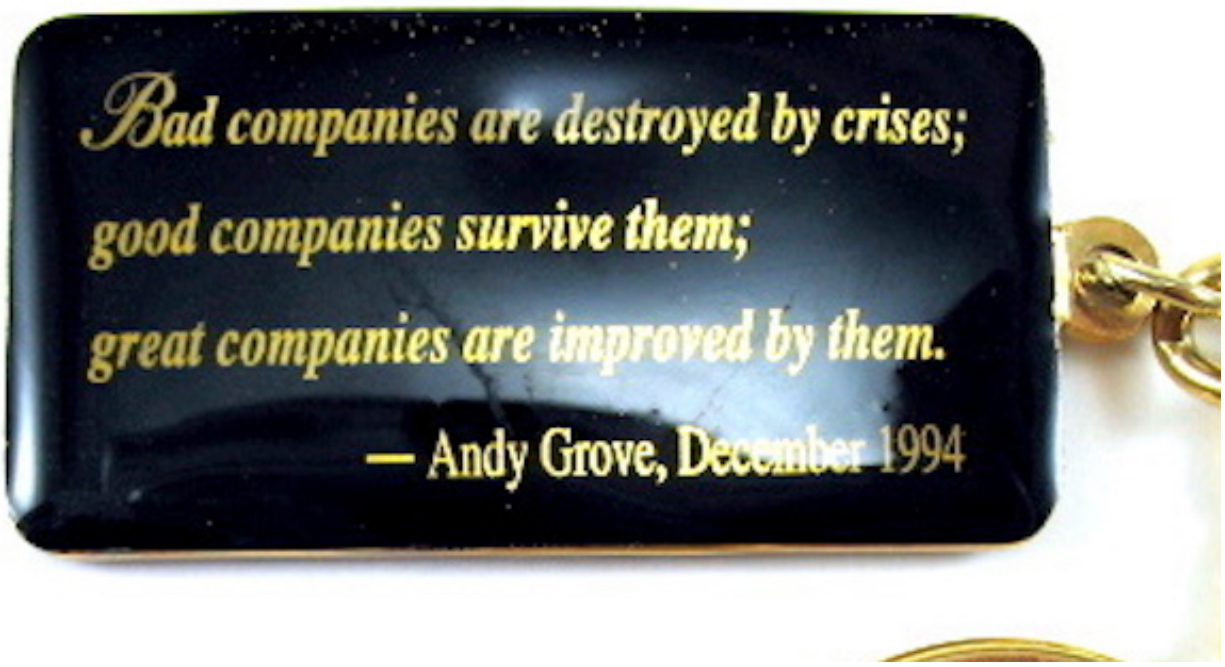
- erreur dans l'algorithme de division (SRT de base 4) ;
- nombreux quotients faux. Pire cas : 4195835.0/3145727.0 donne 1.33373906802 au lieu de 1.3338204491 ;
- tempête électronique sur Internet ;
- Intel a dû remplacer les Pentium défectueux (coût : peut-être 400M$) ;
- la vraie perte a été en termes d'image de marque.

Après ceci : vrai changement de stratégie

- fin du secret sur les algorithmes VF : division de l'Itanium publiée dans les actes d'Arith14 (1999) ;
- preuve formelle : Intel embauche Harrison, AMD embauche Russinoff.

Que sont les Pentium devenus ?

Que sont les Pentium devenus ?



D'autres choses en vrac...

- algorithmes Compensés : pour produits scalaire, évaluation de polynômes, arithmétique complexe, etc.
- arrondi correct des fonctions : dilemme du fabricant de tables (Arénaire, depuis 1998 environ) ;
- coq et flottant (une bonne partie de notre communauté)
- nouvelles moutures (2008 puis 2019) de IEEE 754 :
  - prise en compte de nouveaux opérateurs, de nouveaux algorithmes ;
  - gestion de la co-existence de plusieurs formats ;
  - meilleure insertion de la VF décimale ;
  - nouveautés (recommandation sur fonctions « élémentaires ») ;
  - spécification de formats à « grande précision ».